DOCUMENT RESUME

ED 461 666 TM 033 456

TITLE	Maryland School Performance Assessment Program (MSPAP), 1997. Technical Report.
INSTITUTION	Maryland State Dept. of Education, Baltimore.; CTB / McGraw-Hill, Monterey, CA.; Measurement Inc., Durham, NC.
PUB DATE	1998-11-11
NOTE	95p.
AVAILABLE FROM	For full text: http://marces.org/mdarch/home.htm.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Program Implementation; Reliability; Scoring; *State Programs; Tables (Data); *Test Construction; Test Content; *Testing Programs; Validity
IDENTIFIERS	*Maryland School Performance Assessment Program

ABSTRACT

Maryland School Performance Assessment Program (MSPAP) assessments are criterion-referenced performance tests designed, developed, and implemented by the Maryland State Department of Education in collaboration with classroom teachers and other Maryland educators. MSPAP is the major strategy for implementing Maryland's educational reform initiative. It provides information relevant to assessing school performance and guiding school improvement plans and activities. The primary focus of the information from the MSPAP is schools, although information about individual students is available. In June and July 1997, approximately 170,000 student answer books were scored. This technical report contains information about: (1) MSPAP test forms (clusters) and test groups; (2) test development; (3) scoring; (4) special issues related to mathematics, algorithmic scoring, and student participation in MSPAP; (5) scaling and equating; (6) validity; (7) score interpretation; and (8) the MSPAP score reports. Three appendixes contain test maps, information on the number of items comprising each outcome, and scale score ratings from each MSPAP proficiency level. (Contains 29 tables and 23 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 461 666

TECHNICAL REPORT

1997 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

November 11, 1998

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Moody

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033456

BEST COPY AVAILABLE

INTRODUCTION 4

MSPAP TEST FORMS (CLUSTERS) AND TEST GROUPS 5

TEST DEVELOPMENT 5

Pre-Packaging of Manipulatives 5

Test Administrations 10

SCORING 10

Quality Control During Scoring 12

Conclusion 15

SPECIAL ISSUES 15

Mathematics 15

Algorithmic Scoring 16

Student Participation in MSPAP 16

SCALING AND EQUATING 17

Rater Year Effects Study 22

Equating 1996 and 1997 Scale Scores 23

Coefficient Alphas 27

Standard Errors of Measurement for Proficiency Level Cut Scores 28

VALIDITY 28

Between Content Area Correlations 28

Between Content Area Correlations at the School Level 29

Test Difficulty Concerns 29

Content Validity Evidence 29

Face Validity Evidence 30

Construct Validity 30

Statistical Test Bias 31

Conclusion 33

SCORE INTERPRETATION 33

Scale Scores 33

Proficiency Level Descriptions 33

School Performance Standards 35

Individual Student Scale Scores 37

Outcome Scores 37

MSPAP SCORE REPORTS 39

REFERENCES 40

TABLES 43

APPENDICES 77

TECHNICAL REPORT
1997 Maryland School Performance Assessment Program (MSPAP)

Maryland State Department of Education
CTB McGraw-Hill
Measurement Incorporated

November 11, 1998

Introduction

Maryland School Performance Assessment Program (MSPAP) assessments are criterion referenced performance tests designed, developed, and implemented by the Maryland State Department of Education (MSDE) in collaboration with classroom teachers and other Maryland educators. MSPAP is the major strategy for implementing Maryland's reform initiative and provides information relevant to assessing school performance and guiding school improvement plans and activities. The primary focus of the information provided from MSPAP assessments is *schools*, although information about individual student performance is also available.

Since 1991, MSPAP has been administered to approximately 170,000 students in grades 3, 5, and 8 each May. Each student participates in nine hours of testing (reading, writing, language usage, mathematics, science, and social studies) over a five-day period, approximately one hour and 45 minutes of testing time per day. The assessments are based on the Maryland Learning Outcomes (available from the Maryland State Department of Education) that were adopted by the Maryland State Board of Education in 1990.

MSPAP Test Forms (Clusters) and Test Groups

MSPAP is comprised of three test forms, or clusters, and one equating form or cluster from the previous year's test per grade (e.g., 3A, 3B, 3C, and 3E). Clusters are non-parallel test forms because content areas are matrixed throughout each cluster. For example, in social studies, *Peoples of the Nation and the World, Geography*, and *Economics* might be assessed in one cluster; *Political Systems, Peoples of the Nations and the World*, and *Economics* in another cluster; and *Political Systems, Geography, and Peoples of the Nations and the World* in the third cluster. Each test form or cluster assesses a combination of reading, writing, language usage, science, social studies, mathematics content and mathematics process.

Students are randomly assigned to testing groups. Random testing groups help to ensure that groups of students assigned to take each test cluster are heterogeneous in ability. In addition, random testing groups minimize influences on student performance that may occur when students are assessed in intact classroom groups by their regular classroom teachers.

Test clusters are assigned randomly to testing groups within schools and across schools in each school system and the state. Local Accountability Coordinators (LACs) implement a simple procedure (spiraling) to ensure this random assignment. Spiraling also ensures that the numbers of clusters administered within each school system and across the state will be nearly equivalent, and that schools with only three testing groups will always be assigned each of the three clusters. The Maryland State Department of Education's (MSDE's) Assessment Office approves final cluster assignments.

MSPAP is equated across years through random equivalent groups and equating clusters. Equating clusters are assigned to a representative sample of schools that have four or more testing groups in a grade and that were not used in the previous year's equating sample. Each equating cluster is given a test from the previous year's MSPAP administration so that the current year's test can be adjusted for difficulty.

Test Development

MSPAP assesses school performance on the Maryland Learning Outcomes through assessment tasks--collections of inter-related assessment activities or "items" that are organized around a theme (e.g., *Recycling* or *Salinity*). Tasks require students to respond to questions or directions that lead to a solution of a problem, a recommendation or decision, or an explanation or rationale for the responses. Some tasks assess one content area; other tasks assess multiple content areas. Activities comprising the tasks may be group or individual activities; hands-on, observation, or reading activities; and/or activities

that require extended written responses, limited written responses, lists, charts, graphs, diagrams, webs, and/or drawings.

Test development consists of five phases: planning, design, development, review and revision, and field test followed by further revisions.

Planning. MSDE instructional and assessment staff select tasks from previous MSPAP administrations to be reused. Staff then determine what learning outcomes are needed to complete test clusters and plan new tasks to assess the outcomes. Up to 50% of the test may consist of reused or rolled over tasks.

Design. MSDE instructional staff write task outlines comprised of a topic area, the time allotted for the task, and the outcomes to be assessed. They design calendars showing the types of test activities and the balance of content areas for each day of testing.

Development. Approximately 170 Maryland teachers across grades 3, 5, and 8 are recruited, screened, and hired by MSDE to write MSPAP tasks and activities; develop scoring tools; and write test administration directions. Task writers are given specifications for the content areas and outcomes to be assessed; the numbers of assessment activities per outcome and task; and the background reading materials to be used in the assessment.

Task writers are trained on the principles of performance assessment, characteristics of MSPAP, bias and sensitivity issues, and Maryland Learning Outcomes. They receive information on scoring, measurement, and administration issues; and guidelines for developing graphics and selecting tools and materials. Task writers also receive concentrated training in the areas for which they are responsible: task writing, scoring, or test administration.

Task writers develop drafts of tasks to which reading and writing cues and prompts are added where appropriate. MSDE specialists and task writers participate in an extended review and revision process that includes raising questions and resolving issues and concerns about the tasks.

One characteristic of MSPAP is the use of authentic texts. Local school media specialists select reading materials in topic areas, and reading content area staff review the materials for bias, sensitivity, and readability. After third and fifth grade "average readers" read the materials with the state reading specialist, an analysis is conducted to determine if the readability is appropriate. Only materials that average readers can read independently and show evidence of construction of meaning are used in MSPAP.

Task writers select materials, from the samples provided by media specialists, that can be used in their entirety. Occasionally, the publisher/copyright owner will not grant

permission to use a text or material, and the task must be altered to accommodate other materials. For the 1997 MSPAP, MSDE secured copyright permission for 98 texts and materials.

After tasks have been drafted, they are examined to see that all activities provide a measure of the intended outcomes. Draft scoring tools, answer cue information, and sample responses are then developed. MSDE specialists and staff from the scoring contractor for MSPAP (Measurement Incorporated) review draft scoring tools and test booklets (*Answer Books, Resource Books*, and *Examiner's Manuals*) to identify problems. They then make revisions where necessary.

Review and Revision. MSPAP tasks are reviewed for:

- technical soundness,
- feasibility,
- controversial and sensitive topics,
- developmental appropriateness,
- scorability, and
- clarity.

Assessment specialists conduct technical reviews that include verifying the numbers of outcome measures in a content area and test cluster and the independent responses in a content area. At least eight independent outcome measures for each content area in each cluster are needed for scaling purposes. Four measures for each outcome measured in a cluster are needed to calculate outcome scores. The test design specifies that an outcome be measured in at least two clusters within a grade.

Local Accountability Coordinators (LACs) and assessment staff conduct feasibility reviews that include examining tasks for:

Timing - Is adequate time allotted to tasks? Are the time blocks listed correctly in test materials?

Ease of Administration - Can tasks be administered by all teachers using the same directions?

Setting - Will all classrooms accommodate the administration of each task?

Clarity and Complexity of Directions - Are directions clear and concise?

Cluster Balance - Are content area tasks evenly distributed throughout the week? Is there task variety (e.g., hands on experiment) within a day?

Formatting - Is there adequate student response space in the *Answer Book*?

Tools and Materials - Are materials appropriate? Adequately described? Feasible to administer? Cost effective?

Assessment and content staff conduct controversial and sensitive topic reviews in which they examine tasks for controversial language, stereotyping, and treatment of minorities, genders, and persons with disabilities. To ensure that MSPAP is free from controversial and sensitivity topics, task writers use *Guidelines to Avoid Bias and Sensitivity* that were adapted from *Bias Issues in Test Development* published by the National Evaluation System, Inc.

Third and fifth grade teachers, educational psychologists, and early learning university faculty conduct developmental appropriateness reviews, to ascertain that assessment tasks are developmentally appropriate for the grade level in which they are to be administered.

Assessment specialists conduct scorability reviews to verify that tasks are scorable and that they yield meaningful measures of what students understand and are able to do. Outcome/activity matches, which identify the outcome(s) being assessed by each activity, are verified.

Content specialists conduct clarity reviews to confirm that tasks are clearly written.

After MSPAP tasks have been reviewed, they are organized into an *Answer Book*, a *Resource Book*, and an *Examiner's Manual* for each grade and cluster (3A, 3B, 3C; 5A, 5B, 5C; 8A, 8B, 8C). All test booklets are then reviewed and edited for consistency, accuracy, organization, and comprehension.

Role playing is conducted to ensure that directions and timing are clear and correct. One MSDE specialist is the "teacher" and the other is the "student" who use the *Answer Book*, *Resource Book*, and *Examiner's Manual* as if they were taking the test. This mock administration allows for cross checking of all materials the students and test administrator will need during the actual test administration.

Field Test. A field test is conducted to collect information on the feasibility of conducting tasks in a classroom setting, clarity of directions to students and examiners, reliability of tools and materials, and timing and scorability of tasks.

In October 1996, six schools in Cherokee and Pickens Counties in Georgia administered the 1997 MSPAP field test. The schools were chosen because their student populations closely matched Maryland's population with respect to race/ethnicity, gender, and school achievement. In addition, reading/writing whole language instruction, collaborative learning, and hands-on learning were part of daily instruction. All new tasks appearing on the 1997 assessment were administered to two classrooms, each containing 25 to 30

students.

Observers from Maryland monitored the field test administration process. The responses generated during the field test were used for range finding and the development of scoring tools and guides. As a result of administrative and scoring feedback, some tasks were slightly revised to correct timing, directions, confusing questions, and troublesome tools and materials. After the revisions were made, a post field test meeting confirmed that the test was ready for the May 1997 administration. (Additional information may be obtained from MSDE: Westat, 1997.)

Field test responses also helped to identify possible anchor (rangefinders), training, and qualifying responses for use in scoring training. These sample responses were selected to represent all score points possible and were based on exact agreement after discussion. (Additional sample responses for scorer training were selected from live responses "hijacked" after the MSPAP operational administration in May 1997.)

Development of Scoring Training Materials. Following field test scoring, the scoring contractor reviewed and revised scoring tools, answer cues, and sample responses to create scoring guides for each task. Each activity was presented, followed by the scoring tool and answer cue information (typical response content, key ideas, etc.). Sample responses were selected to illustrate each score point. In the few instances in which field test scoring had not yielded any samples at a given score point, a teacher-developed sample response was utilized. Responses from the May 1997 administration supplemented these teacher-developed samples. Scoring guides were all task-specific, with the exception of language in use. This guide was generic, and was used for anchor responses to a wide array of language usage items.

The scoring contractor's senior staff developed detailed annotations to assist the Maryland-based scoring team coordinators and team leaders to train their teacher teams on scoring MSPAP. In addition, supplementary guides dealing specifically with poetry were developed to assist the expressive writing teams to apply the genre-general rubric to this particular expressive form.

Preparation of Scoring Training Materials. Training materials (training and qualifying sets) were prepared using field test and operational responses. Training sets were used for instruction and practice in task scoring. Qualifying sets were used to test the readers' ability to score accurately and to supplement the training provided by the training sets. These sets included responses from all activities to be scored by the team and were formatted to resemble the portion of the *Answer Book* which the team would score. Work was also begun on the accuracy sets which would be used twice a week during scoring to diagnose and prevent individual and/or room-wide drift away from scoring criteria. These sets closely resembled the qualifying sets described above. Preparation of training materials continued to mid June, when training began.

Pre-Packaging of Manipulatives

Beginning with the 1995 MSPAP, to standardize assessment materials and relieve the procurement burden on local school systems, MSDE contracted with The National Resource Network (a non-profit agency employing handicapped adults) to package materials for hands-on activities for each testing group. Materials are delivered to schools in the 20 school systems participating in the Network. When possible, materials are pre-cut or pre-measured, such as the amount of detergent or soil needed, and packaged for each student or teacher.

Test Administrations

In May, the administration of MSPAP was observed by MSDE content and assessment staff to see how teachers, school staff, and students responded to the tasks, and to gather information on the MSPAP administration. Information gathered as a result of observations was added to examiners' feedback and used during item analysis and during the revision process if tasks had been selected for re-use.

As in other years, test examiners submitted personal comments about the test on the "Concerns or Comments on the Administration of the 1997 MSPAP" form included in the *Examiner's Preparation Guide*. Some examiners made general comments about the test; others commented on specific tasks. Most comments focused on timing (too little or too much) and directions (vague, confusing, or ambiguous).

Examiners' comments are read and collated. Since some tasks will be re-used in the next year's administration, comments on all reused tasks are scrutinized in roundtable discussions. Based on the comments and concerns of the administration, as well as, feedback from other sources, tasks are adjusted as appropriate before they are administered again.

Scoring

Four teams of Maryland teachers scored the assessment activities in each test form at each of the three grades using scoring guides developed by Measurement Incorporated (MI) project staff, scoring tools generated by Maryland educators, and selected sample responses chosen by Maryland educators. Each team scored the open-ended student responses and assigned the appropriate score point on a customized scan sheet. During June and July 1997, *Student Answer Books* for approximately 170,000 students were scored.

The four school sites and scoring assignments for 1997 were:

Clusters 3A and 8A: Mattawoman Middle School, Charles County Public Schools, Waldorf

Cluster 5A: Centreville Middle School, Queen Anne's County Public Schools, Centreville

Clusters 3B, 5B, and 8B: Oakland Mills High School, Howard County Public Schools, Columbia

Cluster 3C, 5C, and 8C: Chesapeake High School, Baltimore County Public Schools, Baltimore

All booklets for a given grade/cluster were scored at the same site due to measurement implications of a multi-site model, as investigated by MSDE staff.

From previous assessments and developmental administrations of various assessment items (e.g., field test), MSDE and MI staff estimated that it would take approximately 25 minutes of reader time to score all scorable units in the answer booklet for each of the 3 clusters at each of the 3 grades—for each of the 9 grade/cluster combinations.

So that scoring loads were reasonable, the scorable units within each of the 9 grade/cluster combinations were distributed across 4 teams. At the eighth grade, a team for each of the four content areas (mathematics, science, social studies, and reading/writing/language usage) scored within their subject areas to the greatest degree possible. Each team scored assessment activities within one primary content area, although content area integration required that teams also address multiple content areas. When integration occurred, enhanced training ensured accurate score decisions by all team members. Additionally, teams were selected to provide a good "fit" with the content areas being addressed by the task(s) being scored by a team. For example, a reading/science task would be predominately scored by a team of science and English/language arts specialists.

At grades 3 and 5, where most teachers work across subject areas, it was not considered crucial that each scoring team score items in only one content area. It was important to attempt to equalize reader scoring time per team, and to ensure that no one team was responsible for too many items requiring mentally demanding, complex thought processes, which might negatively affect the accuracy of readers and teams due to mental fatigue.

Staffing and Reader Distribution Throughout Scoring Sites. For each grade and cluster, four teams scored a unique set of MSPAP items--a total of 12 teams per grade and 36 teams across three grades. For each team, the data processing contractor provided a customized answer sheet. Each student's answer booklet had four customized answer sheets included with it when delivered to the scoring site.

Based upon five years of experience, MI project management established a target of 740 readers to score the 1997 MSPAP assessment, with each reader working about 20 days after 2 to 3 days of training and qualifying. The number of readers required for each team varied depending upon the estimate of the relative scoring time per customized answer sheet after the 36 teams had been created. The average number of readers per team was 21. However, team size varied from 13 to 27 readers distributed across sites, grades and clusters as shown in Table 1.

See Table 1

Each team had two leadership positions. The Scoring Coordinators received five days of training by MI Project Leaders in preparation for training readers on their teams, monitoring readers for quality and production during the scoring process, and administering all aspects of the scoring in concert with MI project staff. The Team Leader received three days of training by the Scoring Coordinator and MI project staff. The Team Leader assisted the Scoring Coordinator in every aspect of the project.

Quality Control During Scoring

After reader training was complete, quality control procedures ensured accurate scoring of student responses. The four major components of quality control were: check sets, accuracy sets, spot checks, and retraining.

Check sets. Check sets for each team using copies of actual student responses were prepared jointly by MI Project Leaders and the Scoring Coordinator and Team Leader of each team. The check sets covered all activities, and multiple score points were represented for activities that generated the most divergent responses. They were administered on Monday morning.

Check sets helped Scoring Coordinators and Team Leaders determine whether individual readers and the team of readers continued to score accurately and consistently, especially on items that were more complex and difficult to score. As reading progresses after training, it is not uncommon for readers to "drift" away from score points, especially for activities requiring holistic decisions, and especially after a weekend away from scoring. The results from check sets were used to "recalibrate" the readers. As inconsistencies and inaccuracies were detected, Scoring Coordinators and Team Leaders held discussions with the team of readers or assisted individual readers to improve accuracy. In addition, individual reader responses to specific items on a check set or a low score on the total check set indicated the need to read behind a particular reader as a spot check to see if retraining was appropriate.

Accuracy sets. Accuracy sets determined whether teams of readers maintained appropriate levels of accuracy during the scoring process. Therefore, each accuracy set included a student response for each scorable unit, and each reader's average score was recorded so that the mean score for each accuracy set could be calculated. These mean scores were used to construct Tables 2 through 7, which will be used to analyze quality control for this scoring project.

Accuracy sets were constructed jointly by MI Project Leaders and the Scoring Coordinator and Team Leader of each team and administered on Tuesday and Thursday mornings. Readers in 35 of the 36 teams were given at least 5 accuracy sets, usually 6 to 7 sets. Readers who scored below 70 percent on any accuracy set received additional training immediately, from the Scoring Coordinator or the Team Leader. The leaders used the results from the accuracy sets to retrain individual readers. Readers were released from the individual retraining process only after the leaders determined that scoring problems were resolved.

Spot checking. A third component of quality control was spot checking, in which a Scoring Coordinator or a Team Leader scored the same booklet scored by a specific reader to estimate a reader's overall accuracy or to determine specific items with which a reader was having difficulty. This general technique is used for routine monitoring of readers who score performance assessment items.

Leaders spot checked readers who had exhibited lower accuracy in recent monitoring. For example, after qualifying was completed and scoring began, the team leaders were to read behind scorers who had the lower results on qualifying rounds, to ensure that they were maintaining accuracy levels and improving with more practice. As the project proceeded, leaders read behind those who had lower results on the most recent check sets and accuracy sets.

Spot checks also helped leaders determine specific items that were causing individual readers to perform poorly on check sets or accuracy sets. If reading behind the individual reader on several student answer booklets pinpointed a limited number of items that were causing scoring problems for the reader, efforts to help readers improve accuracy levels could focus upon the specific items for more efficiency in the retraining process.

Retraining. The fourth component of quality control was retraining, leaders working with individual readers who had problems maintaining appropriate accuracy levels. Retraining involved either the Scoring Coordinator or Team Leader working with an individual reader or a small group of readers who shared a common difficulty in scoring one or more items. The leader used the scoring guide and student papers to help readers score a specific item more accurately. Another technique which often proved effective as the project proceeded was to have one of the more accurate readers work with the readers having difficulty. Sometimes they had a fresh perspective in discussing the score points

and why certain papers should receive a particular score. For some readers who had scoring problems, it was less threatening and more productive to work with a colleague rather than a leader. After it was determined that the scoring difficulties were resolved, the reader continued scoring with the team.

Reader accuracy results. There were 213 accuracy sets administered across all 36 scoring teams in 1997. The reader accuracy set mean scores for each scoring team are shown in Tables 2, 3, and 4 for grades 3, 5, and 8 respectively. The results are summarized in Table 5 by grade and across all three grades.

See Tables 2-5

The results are reasonable and acceptable for scoring open–ended performance assessment items. Fifty-one percent (109 of 213) of the sets had mean scores between 80 to 89% and 36 percent were at or above 90% accuracy. Twenty-four percent had mean set scores between 70 to 79%, and only four of the accuracy set mean scores were below 70% accuracy. The results for the 1997 MSPAP were similar to those for the previous three years. The accuracy set mean scores are similar to past years.

The averages across the accuracy sets for each team could be calculated because the sets contained the same number of scorable units. However, it was not possible to calculate the averages across different teams because the number of scorable units varied considerably from team to team. When the accuracy set mean scores were studied in terms of content area, the results were reasonably predictable yielding no major surprises.

Bearing in mind that few teams addressed only one content area, it is possible to look at results for predominant content area focus in the eighth grade. Results by content area for the eighth grade are displayed in Tables 6 and 7. From past scoring of performance assessments it was reasonably predictable that the scoring of mathematics would yield relatively higher and somewhat more consistent accuracy set scores. As commonly found in the handscoring of performance activities, the accuracy set mean scores for reading/writing/language usage were lower than those for mathematics, science, and social studies.

See Tables 6-7

In grades 3 and 5, the items to be scored within each content area were distributed across teams to such a degree that it was not possible to analyze accuracy set mean scores systematically by content area. Past experience in scoring open–ended performance

assessment items indicated that the relationships between content area and accuracy set scores at grades 3 and 5 would be similar to those at grade 8. In addition, MI Project Leaders and the Scoring Coordinators and Team Leaders felt that it was more difficult to train readers to score items consistently in reading/writing/language usage than in other content areas. These responses more often measure higher level skills and objectives; and they more often require holistic scoring decisions rather than more discrete decisions.

Conclusion

The factors that interacted to produce improvements in training and scoring productivity are:

- Early field testing to provide an adequate time frame for scoring booklets, selecting training materials, and preparing annotated scoring guides.
- An adequate time frame for planning and implementing activities for both CTB, the data processing contractor, and MI.
- Increased experience by project staff at MI and in Maryland. Not only had many readers and leadership staff in Maryland gained another year's experience in scoring MSPAP activities, many of these educators became increasingly involved in other MSPAP activities, such as task development or rangefinding (field-test scoring).

Special Issues

Mathematics

Prior to the 1996 MSPAP, 13 outcomes were measured in mathematics. Having more than twice as many outcomes as the rest of the content areas made designing the mathematics component difficult. The number of measures needed in a cluster often made individual tasks too long. For design reasons, some outcomes were combined to bring the total number to nine. This does not change instruction because all outcomes are still tested, but there needs to be fewer mathematics measures. For example, since geometry and measurement were combined, instead of needing four measures of each outcome for reporting purposes, only four total measures are needed. The mathematics supervisors in each school system accepted this change.

The 1997 MSPAP included limited problem solving. The problem solving outcome has been difficult to include in the test because of the scope of true problem solving. Additionally, scoring time and training needed to be slightly modified. However, it was important to include problem solving activities because of its emphasis at the national and state levels.

Algorithmic Scoring

Algorithmic scoring is a process for deriving a score that uses all available score data in a content area for a student. The process uses a maximum-likelihood estimation which is a general method of finding good parameter estimates in a model. Since table scoring is based on complete score records, the ability estimates of absent students are inaccurate (underestimated). Therefore, students scored algorithmically can have their ability more accurately estimated using a maximum likelihood estimator which approximates student ability using the data available. For the 1996 and 1997 MSPAP, CTB McGraw-Hill scored all students algorithmically. (Before 1996, CTB used table scoring.)

To be eligible for algorithmic scoring, a student must have attempted at least 60% of the content area and at least eight independent items. Exceptions included the content areas of writing and language usage, as well as any "short" test. A short test is a test of fewer than eight independent items. Short tests, typically math process, will not be eligible for algorithmic scoring. Since the mathematics total score is a combination of mathematics content and process,
mathematics does not benefit from this scoring process. Because writing is a three-item test, if a student responds to the extended writing prompt (scored 0-3) and to one of the two limited writing prompts (scored 0-2), then a student should receive a score. (From 1992 to 1994 only one extended and one limited writing process comprised the writing test. Therefore, MSPAP added another limited writing process to the writing scale in 1995. If students missed one of the limited writing process prompts, they still received a writing score.) As identified in the example above, the content area most vulnerable to absence vulnerability is language usage, since language usage measures are captured throughout the week. Therefore, language usage is scored for absent students as long as six or more of the responses in the student's language usage vector have score codes.

Algorithmic scoring increased the number of students who received at least one score. In 1997, across all grades and content areas, more than 15,000 more scores were computed using algorithmic scoring. This method of scoring gave a more accurate reflection of the students within a school or system.

Student Participation in MSPAP

It is the policy of Maryland to include all students to the fullest extent possible in all state assessment programs. Testing accommodations that meet state guidelines are provided to help students with disabilities and English as a Second Language (ESL) students participate more fully in assessments and better demonstrate their knowledge and skills.

MSPAP permits five categories of accommodations (scheduling, setting, equipment, presentation, and response) with 31 accommodations under the five categories for

students with Individualized Education Programs (IEPs) and ESL students. Most accommodations do not invalidate student scores; however, in some cases, the student will not receive a score if the validity of the work that has been accommodated has been compromised. For example, if an examiner must read sections of the test to a student, the reading construct has been comprised. The student is not reading but listening; therefore, the student will not receive a reading score for the test. The student will, however, receive scores in all other content areas.

Students with disabilities may be exempted from MSPAP if they are not pursuing the Maryland Learning Outcomes but, instead, are pursuing alternative or life skill outcomes. ESL students may be exempted if they do not have the minimum language proficiency required for participation in MSPAP. ESL exemptions are limited to one test administration, i.e., a student exempted in grade 3 cannot be exempted again in grade 5.

Students may be excused from testing for a variety of reasons, such as demonstrating inordinate frustration, distress, or disruption of others and/or require accommodations that the school is unable to provide.

Students who are exempted do not take the test and are not included in the calculation of MSPAP scores for a school. Students who are excused do not take the test, but are included in the calculation of MSPAP scores. In other words, the school is not held responsible for students who are exempted from the test; it is held responsible for students who are excused from the tests.

Scaling and Equating

MSPAP is horizontally but not vertically equated. In other words, MSPAP establishes equivalent scores on test forms. The test does not establish equivalent scores across grades (e.g., grades 3 and 5). Therefore, the MSPAP scores can be compared across years within a grade, but not between grades.

Equivalent-Group Design and Analysis: Overview. The equivalent groups design involves administering the tests to be equated to groups of examinees who are equivalent in terms of the skill measured by the tests. In MSPAP, the design is implemented by randomly assigning students to test groups by their Local Education Agency (LEAs). For the cluster equating, at least three test groups of randomly assigned students were created within each grade in a school, and each group was administered one of the three clusters. This procedure resulted in approximately 19,000 students in a given grade assigned to a cluster. For calibration purposes for cluster equating, 7,500 students were randomly selected per cluster within a grade.

For the 1997 annual equating, 2,500 students per grade were selected to take a 1996 cluster. Within each LEA, one or more schools were randomly selected; within each school, a test group composed of randomly assigned students was selected to take the 1996 cluster. Within an LEA, sufficient numbers of schools were chosen so that the LEA's representation in the equating group was proportional to the LEA's representation in the state as a whole.

A caveat must be attached to this description. The pool of schools included only schools that had four or more test groups in a grade because MSDE requires that a minimum of three test groups in a school take the 1997 MSPAP.

The next step in the equating study was to identify a group of students in each grade who took the 1997 MSPAP and who were equivalent to the 1997 group of students administered the 1996 MSPAP cluster. Following MSPAP administration, CTB counted the number of valid students from each LEA who took the 1996 MSPAP for the equating study and randomly sampled from the equating schools in the LEA the same number of students who took the 1997 MSPAP. This procedure ensured that the numbers of students from each LEA were identical in the two groups used for the equating.

The critical assumption that must be met to use the equivalent groups design is that the groups taking the tests to be equated are equivalent, not representative. CTB proportionally samples from all LEAs to construct equating groups to avoid the appearance that any undue influence on the equating results is exerted by one LEA or another.

CTB also performs a rater-year effect equating. In this equating, approximately 1,500 *Student Answer Books* per grade from the 1996 MSPAP administration were rescored by 1997 raters. These data helped to determine and adjust for systematic refinements in rater leniency.

Analysis procedures. The equating process involves constructing an equation that permits the translation of scores obtained on one test to corresponding scores on a second test. It was the responsibility of CTB to express the 1997 obtained MSPAP scores on the 1992 score scale so that performance in the test years are comparable.

The method used derives a linear equation that can be used to adjust the scores on one test so that they correspond to the scores given for comparable performance on the target test. In the case of cluster equating, this target test was the 1997 cluster that had the most regular cumulative score distribution. In the case of the 1996-1997 equating, this target was the 1996 clusters administered in 1997 for the equating study.

When tests are scaled using item response theory, it is necessary that linear equating be done. Traditionally, linear equating based on equivalent groups has involved merely

equating means and standard deviations. However, considering only means and standard deviations can produce unsatisfactory equatings for tests such as MSPAP that have few items or unusual score distributions. Therefore, for equating MSPAP a procedure was used that was more detailed and robust than equating means and standard deviations. This procedure determined the linear transformation that most closely aligned the greatest number of score points possible. This approach is called the linear equipercentile procedure.

The linear equipercentile procedure had several steps. First, an equipercentile procedure identified pairs of scores on the two tests that had the same percentile rank. Then, the linear function was determined that most accurately described this equipercentile result. For the vast majority of tests, the score pairs fell on a straight line; therefore, the linear function ran through all the pairs.

As in previous years, the operating principle for equating was "the greatest accuracy for the greatest number." In other words, the equating line was located so that it passed through as many scores as possible. It was also located with attention on the Proficiency Level 3/4 cut score.

Samples. As in previous years, the calibration of 1997 MSPAP items was done separately by cluster. The calibrations for each cluster were based on stratified random samples drawn from the pool of students in the state who were administered the cluster. The strata consisted of the 24 Maryland LEAs. Within each grade, students were sampled such that their proportional representation in the calibration sample corresponded to their LEA's proportional representation in the state. Table 8 shows that the sample sizes for each calibration ranged from 7,499 students to 7,501 students. Separate samples were drawn for each set of items to be calibrated.

Item Set Calibrations and Analysis of Item Fit. Table 8 shows that item calibrations, or item scalings, were carried out for reading, writing, language usage, mathematics content, mathematics process, science, and social studies. Mathematics content and mathematics process items were assigned to different scales because it was known that some of the mathematics process items would be dependent on the mathematics content responses.

Table 8 shows that no items were deleted due to group administration or at the request of MSDE prior to the initial scaling.

The Two-Parameter Partial Credit model (CTB McGraw-Hill, 1992, p. 4-4), as implemented by the PC based program PARDUX (Burket, 1992), was used for scaling the responses to the 1997 MSPAP items. Trait estimates as well as standard errors of measurement for these estimates were developed using the same procedures that were used in previous test editions. For two items assessing writing content, PARDUX could not provide parameter estimates. These items typically had difficulties that were extreme

and different from the other items in the scale. For each of these items, plots of students' observed performance were used to fit tracelines "by hand." That is, the graphical display capability of PARDUX was used to examine observed item tracelines. Item parameters that produced tracelines that most accurately represented the observed data then were identified interactively.

The same two types of model fit analyses used to evaluate MSPAP items in the past were used again in 1997. The two types of analyses used an analogue to Yen's Q_1 (Yen, 1981) fit statistic and an analogue of Yen's Q_3 dependency statistic (Yen, 1984). The Q_1 statistic was used to compare observed and expected tracelines statistically. Also, graphical representations of these lines were examined. The Q_3 statistic was used to examine local dependence. Even though local dependence is still examined, it is important to remember that there have been no testlets of dependent items constructed since 1992.

Items with differences between students' observed and expected performance that exceeded criterion values were flagged for further study. These criterion values are described in detail in the Technical Report for the 1991 MSPAP. The items that exceeded the criterion values used for the 1997 MSPAP are given in Table 8. Math process and reading had some items flagged for poor fit.

There are limitations to the usefulness of fit statistics such as Q_1. First, chi-square measures such as Q_1 are greatly influenced by the deviation of observations from very small expectations; this influence results in high chi-square values for deviations of no practical significance. Another limitation is that performance on an item is implicitly included in the model via the trait estimate. With shorter tests, such as math process and writing, there is substantial part-whole contamination in comparing item observed performance with predictions that implicitly include that item via that trait estimate. Lastly, the Q_1 statistic criteria is very conservative; it often flags items that in fact fit really well. Due to these limitations, the Q_1 statistic is used as a flag for potential misfit. The fit of each flagged item was then further evaluated using detailed fit information and both graphically within PARDUX.

If very large differences between students' observed and expected performance occurred on an item, the item was judged to have poor fit and was deleted. Table 8 shows that in 1997 no items were deleted due to poor fit.

When reading for literacy experience is measured, students in cluster 3A, 5B, and 8C are allowed to select from three or four passages the one they want to read. When writing for personal expression is being measured, students in 3A, 5B, and 8C are allowed to choose what they want to write about and the form of writing they want to use. Table 9 details the calibration information for the reading and writing choice clusters. For reading choices, the sample sizes ranged from 748 to 3,924. Sample sizes for the writing choices ranged from 380 to 4,673. The writing choices of poem and play are not widely selected

by students. The fit of each flagged item was then further evaluated using detailed fit information and both graphically within PARDUX. Table 9 shows that no items were deleted due to poor fit.

See Tables 8-9

Equating the Content Area Scores Across Clusters. The procedures used to equate the content area scores are comparable to those used to equate the content area scores of previous MSPAP forms. Specifically, cumulative scale score distributions for the calibration sample for each cluster and content area were obtained. In each grade, the content area scores of one cluster were designated as the target distribution. FLUX was used to carry out an equipercentile equating procedure to align distributions of content area scores from each of the two other clusters so that they matched the target distribution as closely as possible. A linear transformation that produced the closest alignment between the target and a non-target score distribution was identified and used to adjust the non-target scores to the score scale.

Table 10 specifies the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) for each content area and cluster. Note that the LOSSes and HOSSes are the same for the three clusters used to assess a given content area in a grade.

See Table 10

Table 10 also indicates the percentage of students in the calibration samples at the LOSS and the HOSS, which is a useful measure of floor and ceiling effects. The table shows that there are substantial floor effects in writing and language usage. These tests are uniformly difficult and very short, and many students in the calibration samples received scale scores at the LOSS.

Linking 1996 and 1997 Scale Scores

The results of two studies were used to express students' performance on the 1997 Maryland School Performance Assessment (MSPAP) on the 1996 score scale. The first, *Rater Year Effects Study*, was designed to determine differences between raters who scored the 1996 MSPAP and raters who scored the 1997 MSPAP. The second, *Equating Study*, was designed to equate the scores of two samples of students who were administered the 1996 and 1997 MSPAPs in 1997.

The results of the two studies were combined to produce values that could be used to transform students' 1997 MSPAP scale scores to the 1996 score scale. This transformation permits comparisons to be made between the performance of students administered the MSPAP in 1996 and 1997.

Rater Year Effects Study

Method. For this study, the responses of approximately 1,500 randomly selected students who had taken the 1996 MSPAP (Clusters 3F, 5F, or 8E) were re-scored by raters who scored the 1997 MSPAP. The 1997 raters were trained, using Scoring Guides developed for the 1996 MSPAP, by Measurement Incorporated (MI), the hand-scoring contractor for the MSPAP in both 1996 and 1997.

Analyses. Analyses of the rater effects were conducted separately by scale within Grades 3, 5, and 8. To determine the magnitude of the rater effect for each scale, the 1996 item parameters were used to generate 1996 scale scores for the students in the study. The first set of scale scores ($96SS_{96}$) was based on the ratings that the students received when they were tested in 1996; the second set ($96SS_{97}$) on the ratings received when they were re-scored by 1997 raters. Both sets of scale scores were expressed on the 1996 score scale.

Linear equipercentile equating procedures, as implemented in the computer software program FLUX (Burket, 1992), were used to align the $96SS_{97}$s with the $96SS_{96}$s. The linear transformation that best expressed the adjustment to the $96SS_{97}$s was used to define the magnitude of the rater effect for each scale assessed in each of the three grades.

Results. Table 11 shows the mean 1996 scale scores ($96SS_{96}$) for the samples used in the Rater Effects Study and the mean scale scores for the state reported in the 1996 Forms Effects Study for Clusters 3F, 5F, and 8E. In the third grade sample, four content areas were slightly lower, one was exactly the same, one was approximately equal (i.e. social studies), and one was slightly higher than the population. In grade 5, two content areas were slightly lower, two were approximately equal (math content and math process), and three were sligh ly higher than the population on the average. The table shows that for grade 8, the samples tended to have slightly higher scale scores than did the population of students who were administered this cluster. Overall, the differences were typically less than one tenth of a standard deviation.

The average raw scores obtained in 1996 and the values obtained when they were re-scored in 1997 are given and compared in Table 12. Positive values, given in the last column of the table, indicate that the 1997 raters graded the students more leniently than did the 1996 raters; that is, they gave the students higher scores on the average. Negative values, in this column, indicate that 1997 raters graded the students more severely than did 1996 raters; they gave the students lower scores on the average.

This table shows that 1997 raters evaluated the samples similarly to 1996 raters. The 1997 raters evaluated all grade 3 content areas slightly more stringently. The differences were less than one-tenth of a standard deviation in all content areas except for language usage. The differences in average raw scores obtained from fifth grade samples indicate that 1997 raters evaluated grade 5 tests more leniently in the content areas of reading, math process, social studies and science, and more severely in writing, language usage and math content. The average raw score differences demonstrate that 1997 raters were slightly more lenient than their 1996 counterparts in evaluating grade 8 tests for all content areas. The largest discrepancy in average raw scores across all three grades was in grade 8, language usage.

Comparisons between the mean differences reported in the current study and those reported for 1992 through 1996 MSPAP are given in Table 13 in terms of standardized mean differences. Positive differences indicate that raters scoring in the year that the study was done were more lenient than raters scoring in the previous test year. Negative differences indicate that raters scoring in the year the study was done were more severe than raters scoring in the previous test year.

Table 13 shows, that in terms of raw scores, rater effects generally were quite small in 1997, ranging from zero- to one-tenths of a standardized mean difference in either direction for all content areas in grade 3; for most of the content areas in grade 5, except reading and science; and for all content areas in grade 8 except language usage. The 1996 and 1997 results indicate that 1996 and 1997 raters were not consistently more lenient or severe than previous study years. The 1997 results indicate small differences between 1996 and 1997 rater groups.

The values of the multiplicative (R_1) and additive (R_2) components of the transformations that best aligned the $96SS_{97}$s with the $96SS_{96}$s are given in the first two columns of Table 14. When applied to the 1996 parameters, these values adjust the 1996 parameter values for the 1997 rater effects. To illustrate the magnitude of the adjustment, the transformation values were applied to a scale score of 500. The value of 500 was chosen because the average 1996 scale score was near 500. Since the values given in Table 14 are expressed in terms of the scale score metric, they will resemble, but not mirror, the raw score results given in Table 2, since raw scores and scale scores have a non-linear relationship.

See Tables 11-14

Equating 1996 and 1997 Scale Scores

Method. For this equating study, equivalent groups of students administered the 1996 and 1997 MSPAP were required, since no anchor items were available to link the tests

administered in the two years. Accordingly, in 1997 approximately 2,500 third grade, fifth grade, and eighth grade students were selected to take 1996 MSPAP test books in May, 1997, while their counterparts were administered the 1997 MSPAP. The third grade students took Cluster 1996 MSPAP Cluster 3A; the fifth grade students took Cluster 5C; and the eighth grade students took Cluster 8B. These are the same books as those that were used for the Rater Effects Study just described.

The test groups in each grade were selected using stratified random selection procedures. Following a priori decisions to involve no more than one test group per school and to use only Maryland schools with four or more test groups in a grade, schools in each LEA were randomly selected to provide test groups for the Equating Study. Schools were selected separately for Grades 3, 5, and 8. The number of schools selected within each LEA was proportional to the LEA's representation in the state. Within each school selected to contribute a test group in a given grade, the test group was randomly selected; since all eligible students in a grade were randomly assigned to test groups, this test group was representative of the students in the school in the grade of interest.

Students' responses to the 1996 test books were scored by the same 1997 raters who were trained to score the 1996 books for the Rater Year Effects Study. For each scale, the students were screened to ensure that they had ratings for all items used to assess that scale in the cluster of interest.
Only those students meeting the screening criteria were used in the analyses for a given scale. For the 2,500 cases administered a 1996 cluster in each grade, Table 15 shows that the screening process left a minimum of 2,362 students per scale for the analyses.

To develop equivalent groups administered the 1997 test, a priori it was decided to select students who had been administered the clusters used as targets in the 1997 cluster equating. The target clusters typically had the most items, therefore the most reliable measurement. The target clusters also typically had smooth score distributions and items with good fit. The target clusters for the cluster equating in reading were 3E, 5D, and 8E; for writing, 3F, 5F, and 8D; for language usage, 3F, 5F, and 8D; for math content, 3E, 5F, and 8F; for math process, 3D, 5E, 8D; for science, 3E, 5E, and 8E; and for social studies, 3F, 5D, and 8D.

The equivalent groups administered the 1997 target clusters in each grade were developed separately for each scale within the grade. To do this, the number of 1997 students selected from each LEA for the analyses was the same as the number of students from that LEA who took the 1996 test books for the Equating Study and had valid scores on the scale. For example, if in the Equating Study, 24 students from LEA #1 took 1996 Cluster 3A and had valid reading scores, to develop an equivalent group for equating 1996 and 1997 reading scales, 24 students from the same LEA who had valid scores on the 1997 target cluster (3E) were randomly selected.

See Table 15

Analyses. The students in the 1997 Equating Study who took the 1996 test books were scored using 1996 item parameters estimated for the items in these books. The use of these parameters ensured that these students' scale scores would be expressed in terms of 1996 scale scores; since these students' responses were scored by 1997 raters, it is useful to designate these scale scores as $95SS_{96}$. The students who took the 1997 test books were scored using the 1997 item parameters estimated for the items in these books, so that these students' scores were expressed in terms of 1997 scale scores. Since these students' responses were scored by the 1997 raters, their scale scores can be designated $96SS_{96}$. In the equating analyses, the lowest and highest obtainable scale scores from the 1996 MSPAP were used so that the scale scores for all students would not fall beyond the range of scale scores obtainable in 1996.

Equating procedures implemented by FLUX (Burket, 1992) were used to align the $96SS_{96}$s with the $95SS_{96}$s. The linear transformation that best aligned the $96SS_{96}$s with the $95SS_{96}$s was used to express the $96SS_{96}$s on the 1996 scale.

Results. The equivalence of the two samples used in the equating is critical for the soundness of the equating. The only data available to measure the equivalence of these samples were the distributions of students across LEAs, which indicated that equating groups matched exactly in terms of the number of students taken from each LEA.

In the paragraphs that follow, comparisons are made between the test performance of the equating samples administered the 1996 books and the state as a whole in 1996. These comparisons are useful for purposes of documentation and general information.

Table 15 describes the sample of students' $95SS_{96}$s and compares these scores to state means estimated for 1996. In examining this table, it is important to keep in mind that the $95SS_{96}$ reflect performance on 1996 items evaluated by 1997 raters, adjusted for the differences between the 1996 and 1997 raters. In other words, these statistics reflect the scores that would have been obtained had 1996 raters been used.

The table shows that the scale scores are relatively similar across the grades when state and sample results are compared. For grade 3, the differences in means are less than one tenth of a standard deviation in all content areas except for science and social studies. For these two content areas, the performance of the 1997 sample on the 1996 MSPAP equating cluster (i.e., 3A) was poorer relative to the statewide 1996 MSPAP performance. For grade 5, the differences in mean scale scores are less than one tenth of a standard deviation across all content areas. For grade 8, the difference in mean scale scores are less than one tenth of a standard deviation across all content areas, except writing, math

content, social studies and science. The 1997 sample performed better on the 1996 equating cluster (i.e., 5C) than the 1996 MSPAP population for writing; however, the reverse pattern is true for math content, social studies and science. Inspection of the case counts by LEA in each grade revealed that the proportions of students from each LEA were similar to the proportion of students that the LEA represents in the state.

The values of the multiplicative (T_1) and additive (T_2) components of the transformations that best aligned the $96SS_{96}$s with the $95SS_{96}$s are given in the first two columns of Table 16. In addition, the result of applying these transformation values to a scale score of 500 are shown in the third and fourth columns of the table to provide a sense of the size and direction of the test effect. Positive values in the fourth column of the table indicate that a scale score of 500 obtained on the 1997 MSPAP was transformed to a score greater than 500 on the 1996 scale. Negative values indicate that a scale score of 500 obtained on the 1997 MSPAP was transformed to a score less than 500 on the 1996 scale.

See Tables 15-16

Comparison of 1996 and 1997 Mean Scores. Table 17 provides data permitting comparisons between the MSPAP performance of the students in 1996 and 1997. Both the 1996 and 1997 results reflect the average scale scores obtained by the student populations in grades 3, 5, and 8.

The results in Table 17 suggest that there was a slight improvement in student performance for some content areas and grades and declines in others. In grade 3, for example, positive mean scale score differences, suggestive of performance improvement, are seen for reading, writing, language usage, and math process. However, negative mean scale score differences are seen in grade 3 for math content, social studies and science. For grade 5, positive scale score differences between 1997 and 1996 scale scores are seen across all content areas. In grade 8, small negative scale score declines are seen for reading, language usage and social studies, no change in scales scores are seen for science. The remaining content area results suggest slight improvements in performance.

Caution must be exercised when interpreting the differences observed in Table 17. This is especially true for writing and math processes results since they were very short tests and had large standard errors. The differences observed in the third column of Table 17 are too small to allow an interpretation of the trend of the performance of the Maryland students by themselves. However, consistently higher scores for the students suggest some degree of growth has occurred in each grade for several content areas.

When considering these results, it is important to remember that different statistics can be used to describe student performance. Average scores are a convenient statistic, but when

distributions are as skewed as many are for the MSPAP, the median may be a better indicator of typical test performance. Reports produced by Maryland summarize performance in terms of proficiency standards; these bands constitute another set of statistics by which performance can be described. The statistic used will affect the results one obtains and the conclusions one draws about growth or declines in performance over years. The average scores reported in Table 17 may not provide the same picture of student performance as that obtained when other statistics are used to describe this performance.

See Table 17

Review and Decision Points for the 1997 Equating. As an equating assurance check, review and decision points were examined for all cluster and annual equatings. MSDE, the National Psychometric Council, and CTB McGraw-Hill reviewed the cluster scaling and equatings, rater year effect equatings, annual equatings, and performance results before each subsequent step of the process was undertaken. Through this process the test characteristic curves and percentile rank correspondences were found to be very acceptable for the 1997 MSPAP equatings.

Reliability

Coefficient Alphas

Coefficient alpha is a reliability measure suitable when items have a variety of score levels (Allen & Yen, 1979). The coefficient alphas based on the calibration sample are reported in Table 18 by grade and cluster. Refer to Table 8 and 9 for the sample sizes and the number of items comprising each scale. The alpha coefficients for each grade and content area are generally around 0.85 except for writing, which is generally around 0.70. Generally, the mathematics process scale has lower alphas than other scales as well. Both the writing test and mathematics process test are short tests, unlike mathematics content and social studies. For example, the writing test is comprised of three items spanning at least two different writing purposes, unlike mathematics which usually has more than 30 items per cluster. (For information pertaining to the number of items comprising a scale, refer to Table 8). The coefficient alphas for each MSPAP test within each cluster are consistent with other constructed response tests (e.g., see KIRIS Accountability Cycle Technical Manual, 1996).

The coefficient alphas obtained in the MSPAP writing assessment are typical of short tests. The MSPAP writing results are similar to the coefficient alphas obtained on the Maryland Writing Test (MWT), a performance assessment comprised of two items. The coefficient alphas for the MWT range from 0.50 to 0.55. Therefore, the reliabilities for the writing portion of the MSPAP are considered acceptable as well.

See Table 18

Standard Errors of Measurement for Proficiency Level Cut Scores.

The standard error of measurement (SEM) is displayed in Tables 19 to 21. These SEMs are for individual scores in each content area. No test provides an exact point estimate. Instead, all scores have some degree of error. The SEM, produced through the Two-Parameter Partial Credit model, is influenced by the amount of information provided by each item and the number of items contributing to a content area. In this way it is similar to coefficient alpha.

The SEMs for the proficiency level cut scores range from 14 to 40 scale score points in the third grade, from 14 to 47 in the fifth grade, and from 9 to 70 in the eighth grade. The SEMs at the HOSSes and LOSSes are larger. The SEMs for the HOSSes range from 20 to 52 scale score points in the third grade, from 23 to 63 in the fifth grade, and from 28 to 70 in the eighth grade. The SEMs for the LOSSes range from 28 to 112 scale score points in the third grade, from 34 to 67 in the fifth grade, and from 30 to 96 in the eighth grade. As can be noted from the tables, SEMs are usually smaller in the middle of the scale distribution (i.e., Proficiency Level 3/4 cut) and larger at the ends (i.e., HOSSes and LOSSes). Because the SEM is a function of item and test information, higher standard errors of measurement are not surprising in writing, language usage, and math process which are all short tests of three to nine items.

See Tables 19 to 21

Validity

MSPAP validity evidence is collected to support and validate intended interpretations and uses of scores from the assessment. Additionally, it is important that MSPAP assesses the skills and knowledge that are documented in the Maryland Learning Outcomes document. The validity evidence described below is organized around these goals.

Between Content Area Correlations

Correlations were calculated to examine the relationships between the content area scale scores at each grade level. The correlations range from 0.55 to 0.84 across all three grades. The relationships can therefore be described as moderate to strong. In Tables 22 through 24, in third grade the largest relationship is between mathematics and science, and the smallest is between writing and reading. For the fifth grade, the largest relationship is between mathematics and science, and the smallest is between writing and reading. In the

eighth grade, the largest relationship is between language usage and writing, and the smallest is between language usage and mathematics. These findings are similar to the moderate to strong correlations found among MSPAP content area scale scores, CTBS/4, and teacher ratings calculated in a special study of the 1991 MSPAP test edition (see CTB McGraw Hill, 1992, Tables 9-8 through 9-10).

See Tables 22 to 24

Between Content Area Correlations at the School Level

Correlations were also calculated to examine the relationships between the content area scale scores at each school. The correlations range from 0.90 to 0.98 across all three grades. The relationships can be described as strong. In Tables 25 through 27, in third grade the largest relationship is between science and social studies, and the smallest is between language usage and mathematics. For the fifth grade, the largest relationship is between science and social studies, and the smallest is between language usage and mathematics. In the eighth grade, the largest relationship is between science and social studies, and the smallest is between reading and mathematics.

See Tables 25 to 27

Test Difficulty Concerns

MSPAP was developed with standards for the year 2000. The test was built around what students are supposed to be learning. Two impacts of test difficulty are (1) the test information function does not overlap well with student scores, and (2) higher standard errors at the lower and upper regions of the distribution. Since 1992, the fit between the test and student achievement has been improving.

Content Validity Evidence

Content validity evidence refers to the degree to which an assessment reflects the content it was designed to assess. The Maryland Learning Outcomes, the basis for learning, instruction, and MSPAP assessment activities, are based on national curriculum standards and learning theories. For example, the reading outcomes are similar to the NAEP reading assessment objectives and based on the reader response theory. Similarly, the writing outcomes are based on long-recognized modes of discourse, and the mathematics outcomes are based on the National Council of Teachers for Mathematics (NCTM). standards for curriculum and evaluation. The science outcomes are based on Project 2061 by the American Association for the Advancement of Science (AAAS). Additionally, the social studies outcomes are underpinned by the work of many groups including the Association of American Geographers, the Commission on History in the Schools, and the

Joint Council on Economic Education. Moreover, the assessment tasks are developed by content area and grade specialists, specifically teachers. Each task development team is given specifications on which outcomes to assess in their task. After tasks are completed, they are reviewed.

A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during the development of tasks, scoring tools, and scoring guides. A task is reviewed by the task writers, test scoring teams, test administration teams, and is field tested. These reviews allow for the opportunity to confirm that the specified outcomes as defined by the Maryland Learning Outcomes document are being assessed.

Outcomes Coverage

Coverage of outcomes by assessment activities is proportionally balanced according to the relative importance of the outcomes at different grade levels. A high degree of match between assessment activities and the outcomes they assess is ensured through multiple reviews during task development and development of scoring tools and guides. All of these reviews allow for the opportunity to confirm that the specified outcomes are indeed being measured-as defined by the Learning Outcomes document. Appendix B presents the Maryland Learning Outcomes and the number of items measuring each outcome by grade and cluster for 1997 MSPAP.

See Appendix B

Face Validity Evidence

Face validity evidence refers to the accuracy with which the test appears to measure what it is supposed to measure. MSPAP has substantive face validity evidence. t is a performance-based assessment that uses authentic and real-life situations as assessment tasks. In addition, reading selections are full-length published works rather than excerpts contrived for use in a test. Furthermore, the test is administered to random groups of students who work in small groups that reflect authentic situations. MSDE content chairs assign tasks to be written for a group of outcomes.

Construct Validity

Construct validity is considered to be the unifying concept for all views and types of evidence of test score validity (see, for example, Messick, 1989, p. 13). One way to assess the construct validity of MSPAP is to compare its results with similar tests. Since MSPAP reflects the NCTM standards and the reader-response model of reading, MSPAP results can be compared to Maryland's NAEP results.

Performance on the 1997 MSPAP indicates that students are not yet proficient in the Maryland Learning Outcomes and that many schools must make significant improvements

in order to meet the state standards. Maryland's fourth grade National Assessment of Educational Progress (NAEP) reading performance show 26% performing at/above the "proficient" level on the 1994 NAEP Trial State Assessment These results are similar to 1997 MSPAP reading results. Where 36.8% of the state's third graders and 35.6% of the fifth graders scored at the satisfactory level or above in reading.

Results from the 1996 NAEP mathematics assessments are not as similar to 1997 MSPAP results. For example, 22% of Maryland fourth graders performed at/above the "proficient" level in the 1996 NAEP mathematics assessment; however, on the 1997 MSPAP, 41.4% of the state's third, and 48.2% of the state's fifth graders scored at the satisfactory level or above in mathematics. In the coming two years, MSDE plans to examine the relationship between MSPAP and NAEP, including a study to equate MSPAP scores to NAEP achievement.

Statistical Test Bias

As a technical term, 'test bias' is not easily defined. A reasonable conceptual approach is to consider a test biased if students of the same degree of attainment in what the test measures receive reliably different scores on the test. A test that fits this definition would then be biased in favor of those who receive the higher scores and against those who receive the lower scores. The difficulty is, in practice, there is no method available to determine whether or not two different students have the same degree of attainment.

In order to overcome the lack of a 'pure' measure of attainment, overall scores on the test are commonly used as the best available measure in order to evaluate 'bias' at the item level. This approach relies on the assumption that bias, if it exists, is presented in some, as opposed to all, the items on the test. Therefore, to the degree that items are identified as biased, it may be true that the test is biased. However, if no items are identified as biased, then it is a reasonable conclusion that test bias is not a threat to test validity.

Differential item functioning (DIF) procedures examine the possibility that non-essential item characteristics may result in misleading poor performance for minority, female, or other defined groups of students. Although the terms item bias and DIF are used interchangeably, DIF does not necessarily imply unfairness. Evidence of DIF is usually considered as a signal to test developers to examine an item more closely to consider whether or not it is defective before using it again.

Items that are biased against groups of students who take the MSPAP, that is, function differently for different student groups diminish construct validity. A measure of DIF generalized from the Linn-Harnisch procedure (1981) is used to flag differentially functioning items. MSDE has studied items flagged for DIF to inform subsequent assessment task development. MSDE examines performance of African-Americans, Asians, and Hispanics in comparison to Caucasians, and examines the performance of females in comparison with males.

During item calibration, the item parameters estimated for the items assessing a given subject area are used to score **all** of the examinees in the calibration sample. The examinees for each target group (e.g., African American) are then sorted into ten equally numerous score categories (deciles). For each item, using the mean attainment estimate for the examinees of the target group in each decile, the predicted and observed examinee success rates are calculated and compared separately in each decile. A positive difference between the observed and predicted values indicates that the target group members in that decile did better than expected. The positive differences are summed to obtain a positive difference value, D+. Similarly, a negative difference indicates that the target group members in that decile did less well then was expected. The negative differences are also summed to obtain a negative difference value, D-. These two sums of differences are summed to obtain an overall difference, D.

DIF was defined in terms of overall differences in performance and in terms of decile group differences. Items for which |D| was greater or equal to 0.10 were flagged as exhibiting DIF or biased. Items for which |D| was less than 0.10 were called unbiased unless D- were less than or equal to –0.10 or D+ was greater or equal to 0.10.

For the 1997 MSPAP, no item was flagged on the basis of the overall D index for any target group. Table 28 presents the number of items for MSPAP 1997 being flagged as exhibiting DIF using the D+ and D- criteria for each target group. It can be seen that no item was flagged for bias either in favor or against African American or Female target groups in any content area at any grade level. While present, the small numbers of flagged items in the Asian and Hispanic groups may be the result of statistical imprecision due to the relative small sizes of these groups in Maryland.

See Table 28

Consequential Validity Evidence.

Since the primary focus of MSPAP is school improvement and performance, the long-term consequence of using MSPAP is expected to be positive: school improvement. The most salient negative consequences of using MSPAP scores (e.g., state approved school improvement plan, possible management by an outside party) will occur for low performing schools. However, these negative consequences are expected to be short-term (i.e., until such schools function successfully), and can be viewed as positive consequences for schools that need help to improve.

Consequences of using MSPAP scores are also expected for students. These consequences could be positive (as schools improve instruction, student learning and performance improve) or negative (if MSPAP does not provide useful information, schools do not improve instruction, and student learning and performance do not

improve). Other consequences of using MSPAP information are also evident. For example, low performance reported in the 1991 MSPAP resulted in complaints that the test was being used for school and teacher "bashing."

Conclusion

The evidence and arguments for construct validity and other technical information about MSPAP provide reasonably strong assurance that MSPAP scores can be validly interpreted for evaluating school performance and guiding school improvement. Similarly, anticipated positive and negative consequences of using MSPAP scores for these purposes provide support for the reasonableness of using the scores for these purposes. Validation of MSPAP score interpretation and use remains an on-going process.

Score Interpretation

Two types of scores are available and relevant to school performance and for use in school improvement planning: scale scores and outcome scores. These two types of MSPAP scores are discussed below. For a more detailed discussions about score interpretation of MSPAP, consult "Score Intepretation Guide" (MSDE, 1996).

Scale Scores

MSPAP was designed to produce scale scores for the content areas of reading, writing, language usage, mathematics, science, and social studies. MSPAP scale scores indicate a school's level of performance in each content area. MSPAP scale scores range, in general, between 350 and 700. The 1992 MSPAP scale scores for all grades and content areas were designed to have a mean of approximately 500 and a standard deviation of approximately 50. Beginning with the 1992 MSPAP, scale scores from the same grade level and content area have the same meaning and are directly comparable from year to year. They are not comparable across grade levels or content areas because of differences in test content and difficulty.

MSPAP scale scores, like other test scale scores, have little intrinsic meaning other than higher scale scores represent higher performance in a content area. It is expected that MSPAP scale scores will acquire further meaning as they are used. Interpretation of the scale scores is aided by proficiency level descriptions. Proficiency level descriptions were developed to help bring meaning to scale scores and to guide interpretation for school performance and improvement.

Proficiency Level Descriptions

The proficiency levels. Proficiency levels and descriptions are intended to inform and guide interpretation of MSPAP scale scores. They describe what students at a particular level generally know and can do in relation to the Maryland Learning Outcomes. The descriptions generally apply to all students at each level rather than to specific students within a level. Individual students whose scale score locates them at a particular

proficiency level may or may not be able to demonstrate all of the knowledge, skills, and processes contained in that proficiency level description.

Proficiency level descriptions for some proficiency levels have not yet been developed because sufficient numbers of items were not located at these levels. This occurred most often at proficiency level 5. As items on future editions of the MSPAP appear at these levels, these descriptions will be developed. In addition, existing descriptions of other proficiency levels will continue to be refined to include information on performance on outcomes not included in the current descriptions.

Listed in Appendix C are the scaled score ranges for each proficiency level in each content area and grade. Detailed proficiency descriptions for each content area and grade appear in Appendix B of the Score Intepretation Guide (MSDE, 1996).

As Appendix C indicates, each proficiency level represents a range of performances and of scale scores. For example, grade 3 reading scale scores lower than 490 indicate Level 5 proficiency, those between 490 and 529 indicate Level 4 proficiency, those between 530 and 579 indicate Level 3 proficiency, and so forth.

MSPAP emphasizes high standards of performance. Since MSPAP scale scores can range as low as 350, there is a wide range of scores in Level 5. Generally speaking, students at Level 5 do not consistently demonstrate Level 4 proficiency. However, they may have provided some responses to assessment activities that, with increased consistency, would have placed them at Level 4.

Development of the proficiency levels. Just as proficiency level descriptions for some proficiency levels have not yet been developed because sufficient numbers of items were not located at these levels, some cut scores have not yet been determined for a similar reason. As items on future editions of the MSPAP appear at and around these levels the remaining cut scores can be developed.

Committees of local and school based educators followed a professional judgment procedure to determine MSPAP cut scores. These committees (a) matched MSPAP items to proficiency level descriptions for Proficiency Levels 1-5, and (b) used the resulting item classifications to establish the location of the cut scores between each proficiency level for MSPAP.

Development of the descriptions. The committee that established the proficiency level cut scores also developed descriptions for each level. For both the establishment and refinement of the descriptions, committees: (a) examine each assessment activity at a proficiency level, the accompanying scoring criteria for each activity, and student responses to each activity; (b) use their professional judgment to determine and list the knowledge, skills, and processes each activity required of students; and (c) synthesize the lists of required knowledge, skills, and processes into descriptions, in Maryland learning

outcomes terms, of what students at each proficiency level know and can do. The committees who defined the cut scores and have gone through the process of creating the description can refine the current proficiency level descriptions by revising and adding to the existing descriptions.

Interpretation and use of the proficiency levels and proficiency level descriptions. Proficiency level descriptions apply generally to any group of students, based on performances by all students and schools in Maryland. The descriptions are not customized specifically for individual students, single schools, or other groups. To reiterate, they describe in general what students at each level know and can do. One approach to school improvement can involve targeting instruction on knowledge and skills at Level 3 for students at Levels 4 and 5.

School Performance Standards

A cornerstone of the Maryland School Performance Program (MSPP) is the process of setting standards of satisfactory and excellent performance levels for schools to meet by 2000.

Development of the school performance standards. Development of the standards for MSPAP followed the same procedures used in establishing the school performance standards for all areas reported in the annual *Maryland School Performance Report.*

> Satisfactory performance denotes a level of performance that is realistic and rigorous for schools, school systems, and the state. It is an acceptable level of performance on a given variable, indicating proficiency in meeting the needs of students.
>
> Excellent performance denotes a level of performance that is highly challenging and clearly exemplary for schools, school systems, and the state. It is a distinguished level of performance on a given variable, indicating outstanding accomplishment in meeting the needs of students (Thorn, Moody, McTighe, Kelly, & Peiffer, 1990, page 7).

Two groups participated in the standards setting process: the Maryland School Performance Standards Committee and the Maryland School Performance Standards Council. In 1992 the Maryland School Performance Standards Committee consisted of 20 members including representatives from 12 local school systems and staff from MSDE. Committee members included teachers, administrators, content area specialists, and assessment specialists. The seventeen member Standards Council represented local education agencies, local boards of education, the state teacher's union, business interests, students, and the state legislature.

> The process of setting standards included several steps. Initially, the Standards Committee recommended a proficiency level to describe satisfactory and excellent performance and the percentage range of students who should score at these levels (i.e., 60% to 80% at the satisfactory level). These recommendations were reviewed by the Standards Council who refined this work to describe satisfactory and excellent performance by proficiency level
> and set a percent of students who should be in each category. These two steps depended on a group decision reached though a convergence process.

The recommendations from the Standards Council were reviewed by the State Board of Education and comments were given through public meetings. Following the public meetings, the MSPAP standards were formally adopted by the State Board of Education.

The Standards Committee recommended level 3 as the proficiency level that describes satisfactory performance and levels 1 and 2 as the proficiency levels that describe excellent performance. Once the ranges for satisfactory and excellent school performance had been established, the recommendations were forwarded to the Standards Council. They were asked to choose a single percentage for each standard for school performance. The Council concurred with the definitions for satisfactory and excellent performance. In addition, the Council recommended 70% for satisfactory and 25% for excellent. For a given school to achieve satisfactory performance in a particular area/grade level, 70% of students must achieve satisfactory performance (level 3 and above). To achieve excellent performance, a school must meet the satisfactory requirement and 25% of these students must achieve excellent performance (level 2 and above). The State goal is that all schools will reach the satisfactory standards by the year 2000.

Interpretation and use of school performance standards for school improvement planning. The score reports produced by MSDE for each school system and school contain numbers and percentages of students at each proficiency level and at satisfactory and excellent standards. School and system staff use these percentages, along with the proficiency level descriptions, to evaluate their school's performance in relation to the Maryland Learning Outcomes. They also use this information to assess their school's progress in reaching standards.

Only those students tested are considered when determining a school's proficiency level, because of the focus on the strengths and weaknesses of the students in the school. Since the school performance standards focus on how well a school is performing on the outcomes, any student who should have been tested is included in the calculation[1]. This includes students who were excused from the MSPAP test administration and students who were absent during the test administration. Therefore, proficiency level percentages may be higher than standards percentages, because the proficiency level percentages are usually based on a smaller number of students.

Individual Student Scale Scores

Scale scores and outcome scores for individual students are not interpretable because each student takes only one-third of the total test. Since the primary focus of MSPAP is school performance rather than individual performance, individual student scores are not to be used for decisions for individual student's performance.

Outcome Scores

Within each of the six content areas assessed on MSPAP, i.e., reading, there are more specific outcomes, i.e., reading to be informed. Outcome scores are based on subsets of items which comprise a content area scale. These scores are the scores that would be expected on an outcome if a student had taken all of the items which measure that outcome. For an outcome score to be reported, at least four measures of the outcome must be present in the test form that the student took. There are two types of outcome scores: Outcome Scores and Outcome Scale Scores.

Outcome Scores. MSPAP outcome scores range from 0 to 100% and are reported for each outcome assessed in each MSPAP content area. They are conceptually analogous to Maryland Functional Testing Program domain scores and can be interpreted like these scores[2]. Outcome scores indicate the proportion of mastery of the knowledge, skills, processes and other requirements that comprise an outcome area. In other words, the MSPAP school outcome score is the average percentage of all score points available on that outcome that a school achieved across all test clusters administered in the school.

Outcome scores are not directly comparable across grades and content areas within a grade, nor are they directly comparable across years because of differences in content and test difficulty. However, they can be compared using information on the relative difficulty of each outcome. Moreover, outcome scores cannot be directly linked to MSPAP proficiency levels.

Interpretation and Use of Outcome Scores. School improvement teams use profiles of a school's Outcome Scores in a content area along with other information about a school, to determine a school's instructional program's relative strengths and weaknesses in each MSPAP content area.

Content area relative difficulty values are reported on Table 29. Relative difficulty refers to the average proportion of the maximum possible score for an outcome across clusters. The relative outcome difficulty index ranges from 0 to 100%. Lower percentages indicate harder outcomes, and conversely, higher percentages indicate easier outcomes. This information is used in conjunction with outcome score averages. An index of relative difficulty was developed because of the desire to compare outcome score averages within each content area to one another.

See Table 29

Outcome Scale Scores. Outcome scale scores are directly comparable across outcomes in the same content area, across years, and to the MSPAP proficiency levels. These scores are expressed on the MSPAP scale score scale and range, as are the content area scale scores, from 350 to 700. Therefore, they can be interpreted in relationship to the underlying score scale and proficiency levels.

MSPAP Score Reports

The four main types of MSPAP score reports are: Maryland School Performance Standards Reports, Proficiency Level and Participation Reports, Outcome Score Reports, and Outcome Scale Score Reports. MSDE provides these reports at the state, school system, and school levels.

MSPAP Standards Reports. These reports provide information relevant to the school performance mission of the Maryland School Performance Program (MSPP). They report percentages of students at satisfactory and excellent levels of performance and indicate whether the standards for satisfactory and excellent school performance have been met. Information on the numbers and percentages of students by grade, content area, race, and gender is available in the MSPAP Disaggregated Standards Report.

MSPAP Proficiency Level and Participation Reports. These reports provide information relevant to the school improvement mission of MSPP. Proficiency level reports for all students in a school, school system, and the state indicate numbers and percentages of test takers at each of the five MSPAP proficiency levels. They also report numbers and percentages of students who completed assessment activities in each MSPAP content area and received a scale score. Also, numbers and percentages of students who were absent, excused, or exempted from the MSPAP test administration are reported. Information on the numbers and percentages of students by grade, content area, race, and sex is also available in the Disaggregated Standards Report.

MSPAP Outcome Score Reports. Outcome Score reports contain the average outcome score, or percentage of mastery of an outcome, for a school, school system, or the state. The total represents the number of students who received questions pertaining to the outcome on their cluster. The Outcome Score Reports also include percentages of students in four outcome score ranges: 0-25, 26-50, 51-75, and 76-100. This information is intended to provide a general idea of the percentage of students who have displayed little or no mastery of the knowledge, skills, and processes required in an outcome (i.e., those in the outcome score range 0-25) and the percentage who have displayed near complete mastery of the outcome (i.e., those in the range 76-100).

MSPAP Outcome Scale Score Reports. The Outcome Scale Score report contain the median outcome scale score for each learning outcome. The median (50th percentile), the interquartile range (25th to 75th percentiles) and the 5th to 95th. Outcome Scale Score reports can be used to compare outcome performance within a content area. Unlike Outcome Scores, Outcome Scale Scores can be compared in a content area because the Outcome Scale Scores have been adjusted for difficulty.

It is important not to over interpret the relationship between Outcome Scale Scores and proficiency levels. Outcome Scale Scores represent performance on activities that

measure only that outcome. In contrast, proficiency levels are established based on all the outcomes in a content area.

References

Allen, M., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) New York: American Council on Education.

Binkley M., Atash, M. N., & Bourque, M. (in press). Standard setting and reporting. In T. Husen and N. Postlethwaite (Eds.), *The International Encyclopedia of Education*, 2nd ed.

Burket, G. R. (1991). *PARDUX, Version 1.4.* Monterey CA: CTB Macmillan/McGraw Hill.

Burket, G. R. (1991). *FLUX Version 1.0.* Monterey, CA: CTB Macmillan/McGraw-Hill.

CTB Macmillan/ McGraw Hill. (1992). *Final technical report: Maryland School Performance Assessment Program, 1991.* (Available from the Maryland State Department of Education, Baltimore, MD.)

Ebel, R. L. (1979). *Essentials of educational measurement*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Kentucky Department of Education. (1996). *KIRIS Accountability Cycle 1 Technical Manual*: Lexington: Author.

Linn, R. L. & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement, 18*, 109-118.

Maryland State Department of Education. (1993). *Technical report: 1992 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1994). *Technical report: 1993 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1995). *Technical report: 1995 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1996). *Technical report: 1996 Maryland School Performance Assessment Program*. Baltimore: Author.

Maryland State Department of Education. (1996). *Test administration and coordination manual, 1996*. Baltimore: Author.

Maryland State Department of Education. (1996). *Score Intepretation Guide, Maryland School Performance Assessment Program - 1996 MSPAP and Beyond, 1996*. Baltimore: Author.

Measurement Incorporated. (1997). 1997 *Maryland School Performance Assessment Program scoring report*. (Available from the Maryland State Department of Education, Baltimore, MD)

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) New York: American Council on Education/ Macmillan.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Thorn, P., Moody, M., McTighe, J., Kelly, N., & Peiffer, R. (1990, April). *Establishing standards for Maryland's School Systems: A systemic approach*. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.

Westat, Inc. (1997). 1997 MSPAP Field Test Report. Available from Maryland State Department of Education, Division of Planning, Results and Information Management.

Westat, Inc. (1994). *Establishing proficiency levels and descriptions for the 1994 MSPAP assessment program*. (Available from the Maryland State Department of Education, Baltimore, MD)

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Tables

Table 1. Numbers of Teams, Readers, and Scoring Leaders by Site, Cluster, and Grade

Site	Grade/ Cluster	Number Of Teams	Target Number of Readers	Number Of Coordinators	Number Of Leaders
Columbia	3B	4	73	4	4
	5B	4	80	4	4
	8B	4	80	4	4
Total		12	233	12	12
Baltimore	3C	4	81	4	4
	5C	4	82	4	4
	8C	4	84	4	4
Total		12	247	12	12
Waldorf	3A	4	87	4	4
	8A	4	92	4	4
Total		8	179	8	8
Centerville	5A	4	82	4	4
GRAND TOTAL		36	741	36	36

Table 2. 1997 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 3

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	AVERAGE
A1	87	92	88	84	84	83	87
A2	86	85	82	84	89	86	85
A3	90	85	96	89	97	96	92
A4	79	80	84	65	83	90	80
B1	82	97	82	91	74	92	87
B2	97	92	98	76	99	96	93
B3	90	90	97	88	84	85	89
B4	77	81	78	77	87	82	80
C1	90	91	82	83	83	84	86
C2	76	88	84	76	79	--	81
C3	88	82	91	96	93	88	90
C4	91	94	94	88	95	91	92

Table 3. 1997 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 5

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	AVERAGE
A1	85	89	93	80	97	94	--	90
A2	87	83	93	83	84	83	--	86
A3	78	81	84	81	82	--	--	82
A4	66	69	80	80	95	87	--	80
B1	82	88	77	90	83	--	--	84
B2	89	90	93	93	86	94	--	91
B3	84	92	83	87	82	87	--	86
B4	79	89	95	83	91	--	--	87
C1	89	89	91	84	94	95	--	90
C2	85	96	90	91	90	90	--	90
C3	92	93	88	90	--	--	--	91
C4	82	73	81	99	93	90	96	85

Table 4. 1997 READER ACCURACY SET MEAN SCORES BY TEAM
GRADE 8

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	AVERAGE
A1	88	87	89	89	93	88	--	89
A2	78	75	78	79	76	80	--	78
A3	86	83	82	87	80	--	--	84
A4	82	86	89	88	80	91	--	86
B1	94	94	97	93	95	95	93	94
B2	65	71	80	75	75	77	--	74
B3	87	87	93	89	79	84	--	87
B4	92	91	92	90	83	87	88	89
C1	88	92	94	93	95	95	--	93
C2	87	84	86	87	89	88	--	87
C3	86	86	89	89	87	84	--	87
C4	86	83	88	79	94	84	--	86

Table 5. 1997 FREQUENCY OF ACCURACY SET MEAN SCORES BY GRADE

Grade	Less than 70 Percent	70-79 Percent	80-89 Percent	90-100 Percent
3	(1%)1	9 (3%)	33 (46%)	28 (39%)
5	2(3%)	4 (6%)	34 (50%)	28 (41%)
8	1 (1%)	11 (15%)	41 (56%)	20 (27%)
ALL GRADES	4 (2%)	24 (11%)	109 (51%)	76 (36%)

Table 6. 1997 READER ACCURACY SET MEAN SCORES BY CONTENT AREA
GRADE 8

TEAM	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	AVERAGE
Matematics								
A1	88	87	89	89	93	88	--	89
B1	94	94	97	93	95	95	93	94
C1	88	92	94	93	95	95	--	93
Social studies								
A2	78	75	78	79	76	80	--	78
B2	65	71	80	75	75	77	--	74
C2	87	84	86	87	89	88	--	87
Science								
A3	86	83	82	87	80	--	--	84
B3	87	87	93	89	79	84	--	87
C3	86	86	89	89	87	84	--	87
Writing								
A4	82	86	89	88	80	91	--	86
B4	92	91	92	90	83	87	88	89
C4	86	83	88	79	94	84	--	86

*Note: Content areas are somewhat integrated.

Table7. 1997 FREQUENCY OF ACCURACY SET MEAN SCORES BY CONTENT AREA GRADE 8

Content	Less than 70 Percent	70-79 Percent	80-89 Percent	90-100 Percent
Mathematics	0 (0%)	0 (0%)	6 (32%)	13 (68%)
Social Studies	1 (6%)	9 (50%)	8 (44%)	0 (0%)
Science	0 (0%)	1 (6%)	15 (88%)	1 (6%)
Writing	0 (0%)	1 (5%)	12 (63%)	6 (32%)
All Content Areas	1 (1%)	11 (15%)	41 (56%)	20 (27%)

Table 8. Summary Findings from Calibrations

Content Area/ Cluster	Sample Size	No. of Items[1]	No. of Items Deleted[2] GA	MSDE	Fit	No. Items with Hand-Estimated Parameters	No. of Items with Fit > Criterion[3]	No. of Students at Min./Max.
Reading								
3A*	7,499*	24*	0	0	0	0	0	261
3B	7,499	12	0	0	0	0	0	224
3C	7,499	12	0	0	0	0	0	345
5A	7,500	12	0	0	0	0	0	121
5B*	7,500	30*	0	0	0	0	1	126
5C	7,500	12	0	0	0	0	1	122
8A	7,501	12	0	0	0	0	0	303
8B	7,501	13	0	0	0	0	1	298
8C*	7,501	34*	0	0	0	0	12	377
Writing/Language Usage								
3A*	7,499*	14*	0	0	0	1	0	943
3B	7,499	12	0	0	0	0	1	948
3C	7,499	12	0	0	0	0	1	577
5A	7,500	11	0	0	0	0	3	403
5B*	7,500	18*	0	0	0	0	7	492
5C	7,500	12	0	0	0	0	0	858
8A	7,501	11	0	0	0	1	2	680
8B	7,501	10	0	0	0	0	5	605
8C*	7,501	22*	0	0	0	1	3	571
Math Content								
3A	7,499	32	0	0	0	0	2	126
3B	7,499	29	0	0	0	0	1	160
3C	7,499	28	0	0	0	0	0	81
5A	7,500	16	0	0	0	0	0	295
5B	7,500	27	0	0	0	0	0	160
5C	7,500	25	0	0	0	0	3	81
8A	7,501	34	0	0	0	0	3	321
8B	7,501	20	0	0	0	0	2	367
8C	7,501	22	0	0	0	0	3	403

(table 8 continue)

Content Area/ Cluster	Sample Size	No. of Items[1]	No. of Items Deleted[2] GA	MSDE	Fit	No. Items with Hand-Estimated Parameters	No. of Items with Fit > Criterion[3]	No. of Students at Min./Max.
Math Process								
3A	7,499	16	0	0	0	0	1	299
3B	7,499	13	0	0	0	0	7	531
3C	7,499	15	0	0	0	0	1	213
5A	7,500	11	0	0	0	0	1	427
5B	7,500	11	0	0	0	0	3	254
5C	7,500	9	0	0	0	0	4	251
8A	7,501	11	0	0	0	0	3	884
8B	7,501	8	0	0	0	0	6	1180
8C	7,501	8	0	0	0	0	7	405
Science								
3A	7,499	15	0	0	0	0	0	276
3B	7,499	16	0	0	0	0	0	231
3C	7,499	19	0	0	0	0	0	119
5A	7,500	17	0	0	0	0	1	192
5B	7,500	22	0	0	0	0	0	231
5C	7,500	22	0	0	0	0	0	119
8A	7,501	17	0	0	0	0	1	548
8B	7,501	28	0	0	0	0	1	190
8C	7,501	16	0	0	0	0	0	277
Social Studies								
3A	7,499	19	0	0	0	0	0	301
3B	7,499	16	0	0	0	0	0	615
3C	7,499	20	0	0	0	0	0	153
5A	7,500	18	0	0	0	0	0	145
5B	7,500	15	0	0	0	0	0	133
5C	7,500	22	0	0	0	0	0	78
8A	7,501	17	0	0	0	0	1	283
8B	7,501	19	0	0	0	0	0	234
8C	7,501	18	0	0	0	0	2	288

(table 8 continue)

BEST COPY AVAILABLE

[1] No. of items refers to the number of items defined as assessing each content area prior to scaling and before items were deleted for the reasons specified in the next column. For the Reading and Writing/Language Usage items in 3A, 5B, and 8C, the No. of items is the total number of items in all choice sets; students administered these clusters actually responded to fewer items than the total given.

[2] The reasons for the item deletion are designated as GA signifying group-administration; MSDE signifying a deletion requested by MSDE; and Fit signifying poor fit.

[3] The cut-off Z values used for various N counts are as follows:

N	Z >
1,500	4
2,000	5
3,000	8
4,000	11
5,000	13
6,000	16
7,000	19

* This is a choice cluster. Sample size, the numbers of items, and the number of misfitting items for this cluster varied over the choice sets.

Table 9. Detailed Findings from Calibration for Clusters with Choice Sets in Reading and Writing

Content Area/Cluster & Choice	Sample Size	Number of Items	Number of Items with Fit Exceeding Criterion
Reading			
3A	7499	6	0
A	2189	6	0
B	3128	6	0
C	2182	6	0
5B	7500	6	0
A	1592	6	1
B	2020	6	0
C	1758	6	0
D	2130	6	0
8C	7501	7	0
A	3924	7	4
B	1303	7	2
C	1526	7	2
D	748	7	0
Writing			
3A	7499	2	0
Story	4673	1	0
Poem	2446	1	0
Play	380	1	0
5B	7500	2	0
Story	2073	1	0
Poem	2050	1	1
Play	896	1	1
8C	7501	2	0
Story	4203	1	1
Poem	2226	1	0
Play	431	1	0
Other	740	1	0

Table 10. Cluster Equating Results

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Reading				
3A	400	650	7	1
3B(T)*	400	650	6	1
3C	400	650	8	0
5A(T)*	375	675	3	0
5B	375	675	3	0
5C	375	675	3	0
8A(T)*	375	650	4	1
8B	375	650	4	1
8C	375	650	5	1
Writing				
3A(T)*	455	635	22	1
3B	455	635	25	1
3C	455	635	25	2
5A	440	595	20	6
5B	440	595	23	8
5C(T)*	440	595	19	9
8A(T)*	425	625	24	5
8B	425	625	30	4
8C	425	625	18	8

(table 10 continue)

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Language Usage				
3A	450	625	22	1
3B(T)*	450	625	23	1
3C	450	625	22	1
5A	425	625	12	2
5B	425	625	12	2
5C(T)*	425	625	17	1
8A(T)*	425	625	12	2
8B	425	625	13	4
8C	425	625	13	3
Math Content				
3A(T)*	375	650	4	0
3B	375	650	4	0
3C	375	650	2	0
5A	400	650	6	1
5B(T)*	400	650	6	0
5C	400	650	5	0
8A	400	650	8	0
8B(T)*	400	650	8	0
8C	400	650	7	0
Math Process				
3A(T)*	375	650	6	0
3B	375	650	7	0
3C	375	650	3	0
5A	400	650	9	1
5B	400	650	9	0
5C(T)*	400	650	10	0
8A	400	650	12	1
8B	400	650	16	0
8C(T)*	400	650	8	0

(table 10 continue)

Content Area/ Cluster	LOSS	HOSS	% at LOSS	% at HOSS
Social Studies				
3A	400	625	7	0
3B	400	625	8	0
3C(T)*	400	625	7	0
5A	400	625	6	0
5B(T)*	400	625	8	0
5C	400	625	7	0
8A	375	650	6	0
8B(T)*	375	650	5	0
8C	375	650	3	0
Science				
3A	375	650	4	0
3B	375	650	5	0
3C(T)*	375	650	3	0
5A	375	650	5	0
5B	375	650	5	0
5C(T)*	375	650	3	0
8A	375	650	7	0
8B(T)*	375	650	4	0
8C	375	650	5	0

* Target cluster

Table 11. Rater Year Effects Study Performance ($96SS_{96}$) of State and Sample on 1996 MSPAP

		State[1]			Sample		
Grade	Scale	Mean	SD	N	Mean	SD	N
3	Reading	511.2	46.6	18,846	511.2	47.5	1,492
	Writing	519.9	49.2	19,221	519.2	48.6	1,492
	Language Usage	516.8	52.0	19,386	517.9	51.4	1,492
	Math Content	514.2	56.0	19,067	513.2	55.5	1,491
	Math Process	515.3	47.4	19,067	514.6	48.8	1,491
	Social Studies	498.7	45.3	19,189	498.5	46.0	1,492
	Science	507.3	51.2	18,759	506.6	50.9	1,491
5	Reading	509.3	47.2	18,577	508.2	47.4	1,510
	Writing	506.6	56.8	18,956	506.0	56.7	1,511
	Language Usage	522.1	58.6	19,052	523.2	58.0	1,511
	Math Content	518.2	51.9	17,741	518.6	51.1	1,507
	Math Process	509.0	54.8	18,968	508.7	53.2	1,507
	Social Studies	514.6	54.4	19,134	516.2	54.7	1,508
	Science	515.2	53.1	18,802	516.1	53.2	1,501
8	Reading	511.0	37.1	17,437	511.4	37.7	1,500
	Writing	499.4	53.5	17,301	510.0	52.2	1,505
	Language Usage	511.7	54.9	17,585	513.6	53.9	1,505
	Math Content	519.4	46.3	17,375	520.9	44.7	1,505
	Math Process	510.6	58.8	17,395	512.8	57.6	1,505
	Social Studies	511.3	47.8	17,390	512.5	47.4	1,500
	Science	524.9	48.1	17,683	525.9	48.5	1,505

[1] State performance results were drawn from the Forms Effect Study carried out for the 1996 MSPAP. The values reported refer to performance on Clusters 3F, 5F, and 8E.

Table 12. Rater Year Effects Study Raw Score Comparisons

Grade	Scale	N	Raters Used: 1996 Mean	1996 SD	1997 Mean	1997 SD	Mean Diff. (97 - 96)
3	Reading	1492	9.91	5.22	9.53	5.13	-0.38
	Writing	1492	2.77	1.83	2.67	1.81	-0.10
	Language Usage	1492	7.94	5.53	7.36	5.31	-0.58
	Math Content	1491	11.64	6.14	11.63	5.98	-0.01
	Math Process	1491	3.97	2.40	3.79	2.23	-0.18
	Social Studies	1492	14.83	7.58	14.79	7.61	-0.04
	Science	1491	11.68	5.26	11.35	5.04	-0.33
5	Reading	1510	7.52	4.37	9.34	5.11	+1.82
	Writing	1511	2.51	2.03	2.17	1.89	-0.34
	Language Usage	1511	7.50	6.11	7.38	6.06	-0.12
	Math Content	1507	15.53	7.40	15.47	7.40	-0.06
	Math Process	1507	4.87	3.33	5.06	3.41	+0.19
	Social Studies	1508	11.91	6.86	13.23	7.39	+1.32
	Science	1510	17.88	7.46	19.73	8.12	+1.85
8	Reading	1500	10.62	5.08	11.48	5.68	+0.86
	Writing	1505	2.25	1.85	2.59	1.80	+0.34
	Language Usage	1505	6.63	5.34	8.56	5.54	+1.93
	Math Content	1505	10.31	7.52	10.79	7.67	+0.48
	Math Process	1505	5.30	4.36	5.52	4.34	+0.22
	Social Studies	1500	12.35	6.13	12.78	6.48	+0.43
	Science	1505	15.87	8.17	16.75	8.51	+0.88

Table 13. 1992, 1993, 1994, 1995, 1996, and 1997 Rater Year Effects Studies: Comparison of Results in Terms of Standardized Raw Score Mean Differences[1]

		Rater Effects Study					
Grade	Scale	1992	1993	1994	1995	1996	1997
3	Reading	0.0	-0.2	0.0	0.0	0.0	-0.1
	Writing	-0.2	-0.2	0.0	0.2	0.0	0.0
	Language Usage	-0.2	-0.4	0.0	0.2	0.0	-0.1
	Math Content	0.1	0.0	-0.1	0.1	0.0	0.0
	Math Process	0.2	0.0	0.0	0.1	0.0	-0.1
	Social Studies[2]	---	-0.6	-0.1	0.1	0.1	0.0
	Science[2]	---	-0.2	0.1	0.0	0.0	0.0
5	Reading	0.3	-0.2	0.1	0.1	-0.1	0.3
	Writing	0.4	-0.1	0.0	0.1	0.0	-0.1
	Language Usage	0.3	-0.2	0.0	-0.1	0.0	0.0
	Math Content	0.1	-0.1	0.1	0.0	0.0	0.0
	Math Process	0.2	0.0	0.1	0.0	0.0	0.0
	Social Studies[2]	---	0.1	0.2	0.0	0.1	0.1
	Science[2]	---	-0.1	0.1	0.1	0.0	0.2
8	Reading	0.0	0.1	-0.1	-0.2	-0.1	0.1
	Writing	0.0	0.0	0.2	-0.1	0.1	0.1
	Language Usage	-0.2	0.1	-0.1	0.0	0.1	0.3
	Math Content	0.1	-0.1	0.0	-0.1	0.0	0.0
	Math Process	0.1	-0.1	-0.1	-0.1	-0.1	0.0
	Social Studies[2]	---	-0.1	-0.1	0.0	-0.1	0.0
	Science[2]	---	-0.2	0.0	-0.2	0.0	0.1

[1] These differences were obtained by dividing the difference between the current and prior year mean ratings by the square root of the pooled variances of these ratings.
[2] This subject was not assessed in this grade in 1991, so comparisons involving 1991 ratings are not available.

Table 14. Rater Year Effects Study Transformation Values

Grade	Scale	Multiplier R_1	Addend R_2	(A) $(R_1*500)+R_2$	(A) - 500[1]
3	Reading	1.005	0.801	503.301	3
	Writing	1.027	-11.487	502.013	2
	Language Usage	1.035	-12.524	504.976	5
	Math Content	1.026	-12.126	500.874	1
	Math Process	1.064	-30.992	501.008	1
	Social Studies	1.014	-6.913	500.087	0
	Science	1.049	-21.840	502.660	3
5	Reading	0.903	37.384	488.884	-11
	Writing	1.065	-23.613	528.887	29
	Language Usage	1.002	0.347	501.347	1
	Math Content	1.020	-10.592	499.408	-1
	Math Process	1.020	-13.851	496.149	-4
	Social Studies	0.974	2.885	489.885	-10
	Science	0.932	22.680	488.680	-11
8	Reading	0.890	50.715	495.715	-4
	Writing	1.059	-40.594	488.906	-11
	Language Usage	1.104	-76.292	475.708	-24
	Math Content	0.993	1.652	498.152	-2
	Math Process	1.012	-8.971	497.029	-3
	Social Studies	0.939	28.937	498.437	-2
	Science	0.983	4.468	495.968	-4

[1] Numbers in this column were purposely rounded to improve their comprehensibility.

Table 15. Performance of State on 1996 MSPAP and 1997 Equating Sample on 1996 MSPAP

		State[1] ($96SS_{96}$)			Sample ($96SS_{97}$)		
Grade	Scale	Mean	SD	N	Mean	SD	N
3	Reading	511.2	46.6	18,846	512.7	47.0	2456
	Writing	519.9	49.2	19,221	521.9	50.3	2456
	Language Usage	516.8	52.0	19,386	522.8	53.9	2456
	Math Content	514.2	56.0	19,067	513.5	56.9	2455
	Math Process	515.3	47.4	19,067	517.5	49.2	2455
	Social Studies	498.7	45.3	19,189	503.9	46.3	2456
	Science	507.3	51.2	18,759	508.4	53.5	2455
5	Reading	509.3	47.2	18,577	514.3	47.7	2446
	Writing	506.6	56.8	18,956	508.9	57.1	2446
	Language Usage	522.1	58.6	19,052	526.6	57.8	2446
	Math Content	518.2	51.9	17,741	521.6	55.8	2446
	Math Process	509.0	54.8	18,968	512.8	54.4	2446
	Social Studies	514.6	54.4	19,134	518.2	54.5	2445
	Science	515.2	53.1	18,802	518.1	54.0	2445
8	Reading	511.0	37.1	17,437	507.9	41.9	2535
	Writing	499.4	53.5	17,301	501.3	54.0	2535
	Language Usage	511.7	54.9	17,585	509.6	58.1	2535
	Math Content	519.4	46.3	17,375	520.2	48.7	2533
	Math Process	510.6	58.8	17,395	511.5	60.9	2533
	Social Studies	511.3	47.8	17,390	513.2	52.5	2535
	Science	524.9	48.1	17,683	523.9	52.3	2533

[1] State performance results were drawn from the Forms Effect Study carried out for the 1996 MSPAP. The values reported refer to performance on Clusters 3F, 5F, and 8E.

Table 16. Equating Study Transformation Values

Grade	Scale	Multiplier T_1	Addend T_2	(A) $(T_1*500)+T_2$	(A) - 500[1]
3	Reading	0.830	99.292	514.292	14
	Writing	0.860	92.604	522.604	23
	Language Usage	1.375	-164.674	522.826	23
	Math Content	1.147	-52.913	520.587	21
	Math Process	0.701	171.941	522.441	22
	Social Studies	0.970	19.860	504.860	5
	Science	1.040	-9.408	510.592	11
5	Reading	0.786	121.537	514.537	15
	Writing	1.237	-108.160	510.340	10
	Language Usage	1.115	-25.180	532.320	32
	Math Content	1.058	-4.644	524.356	24
	Math Process	0.923	55.241	516.741	17
	Social Studies	1.115	-34.536	522.964	23
	Science	1.068	-15.111	518.889	19
8	Reading	0.698	159.509	508.509	9
	Writing	0.972	13.073	499.073	-1
	Language Usage	1.177	-81.008	507.472	7
	Math Content	0.864	93.371	525.371	25
	Math Process	0.976	29.071	517.071	17
	Social Studies	1.024	7.720	519.72	20
	Science	1.057	-3.822	524.678	25

[1] Numbers in this column were purposely rounded to improve their comprehensibility.

Table 17. Comparison of 1996 and 1997 MSPAP Performance by Grade and Scale

Grade	Scale	1996 State Means	1997 State Means	97 - 96 Difference
3	Reading	510.4	513.9	+3.5
	Writing	520.2	521.6	+1.4
	Language Usage	515.8	524.3	+8.5
	Math Content	514.0	516.1	+2.1
	Math Process	514.2	516.9	+2.7
	Total Math	514.5	516.8	+2.3
	Social Studies	497.4	503.1	+5.7
	Science	507.0	508.6	+1.6
5	Reading	509.0	513.7	+4.7
	Writing	506.5	506.9	+0.4
	Language Usage	522.1	523.4	+1.3
	Math Content	517.6	518.5	+0.9
	Math Process	509.2	511.7	+2.5
	Total Math	513.9	515.5	+1.6
	Social Studies	514.9	516.7	+1.8
	Science	513.4	514.7	+1.3
8	Reading	508.5	510.9	+2.4
	Writing	501.0	502.8	+1.8
	Language Usage	511.5	510.2	-1.3
	Math Content	520.5	521.0	+0.5
	Math Process	513.1	513.0	-0.1
	Total Math	517.7	517.2	-0.5
	Social Studies	510.3	516.9	+6.6
	Science	524.1	525.2	+1.1

Table 18. Coefficient Alpha for 1997 MSPAP Content Areas

Grade 3

	Cluster		
	A	B	C
Reading	.86	.83	.82
Writing	.70	.72	.72
Language Usage	.90	.93	.91
Math Total	.90	.89	.88
Math Content	.89	.89	.87
Math Process	.80	.82	.78
Science	.83	.85	.85
Social Studies	.85	.85	.86

Grade 5

	Cluster		
	A	B	C
Reading	.82	.84	.82
Writing	.61	.61	.73
Language Usage	.89	.90	.91
Math Total	.87	.90	.90
Math Content	.83	.89	.88
Math Process	.78	.75	.70
Science	.84	.86	.85
Social Studies	.84	.83	.85

Grade 8

	Cluster		
	A	B	C
Reading	.86	.88	.90
Writing	.74	.73	.77
Language Usage	.92	.92	.92
Math Total	.91	.88	.88
Math Content	.91	.87	.88
Math Process	.79	.77	.73
Science	.88	.90	.84
Social Studies	.87	.88	.89

Note: Clusters 3A, 5B, and 8C are choice clusters.
The reported alpha for the choice cluster are the average alpha across all choices.

Table 19. **Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 3**

		Cluster		
Reading	**Scale Score**	**3A**	**3B**	**3C**
SE at HOSS	650	47	40	35
SE at Level 1/2	620	31	27	22
SE at Level 2/3	580	21	18	17
SE at Level 3/4	530	15	16	15
SE at Level 4/5	490	15	17	17
SE at LOSS	400	36	37	50
Writing				
SE at HOSS	635	36	35	42
SE at Level 1/2	614	30	32	32
SE at Level 2/3	577	26	26	27
SE at Level 3/4	528	27	26	26
SE at LOSS	455	55	49	36
Language Usage				
SE at HOSS	625	21	20	22
SE at Level 1/2	620	20	18	22
SE at Level 2/3	576	20	16	18
SE at Level 3/4	521	21	17	19
SE at LOSS	450	39	34	28
Math Content				
SE at HOSS	650	22	29	30
SE at Level 1/2	626	19	21	24
SE at Level 2/3	583	16	15	19
SE at Level 3/4	531	16	14	17
SE at Level 4/5	489	18	18	20
SE at LOSS	375	43	45	35
Math Process				
SE at HOSS	650	26	52	35
SE at Level 1/2	626	19	28	22
SE at Level 2/3	583	14	16	16
SE at Level 3/4	531	14	12	16
SE at Level 4/5	489	17	20	21
SE at LOSS	375	86	112	80
Science				
SE at HOSS	650	32	35	34
SE at Level 1/2	619	23	25	26
SE at Level 2/3	580	19	20	21
SE at Level 3/4	527	18	17	18
SE at Level 4/5	488	20	40	19
SE at LOSS	375	57	46	33
Social Studies				
SE at HOSS	625	28	23	24
SE at Level 1/2	622	28	23	23
SE at Level 2/3	580	19	15	18
SE at Level 3/4	525	15	15	15
SE at Level 4/5	495	16	18	16
SE at LOSS	400	35	52	30

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

Table 20. **Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 5**

		Cluster		
Reading	**Scale Score**	**5A**	**5B**	**5C**
SE at HOSS	675	60	54	63
SE at Level 1/2	620	31	30	35
SE at Level 2/3	580	22	20	22
SE at Level 3/4	530	16	16	17
SE at Level 4/5	490	16	15	16
SE at LOSS	375	44	50	39
Writing				
SE at HOSS	595	47	47	41
SE at Level 2/3	567	45	44	36
SE at Level 3/4	522	45	44	36
SE at Level 4/5	488	49	47	41
SE at LOSS	440	54	54	58
Language Usage				
SE at HOSS	625	24	26	21
SE at Level 1/2	597	17	18	16
SE at Level 2/3	567	15	13	13
SE at Level 3/4	533	17	14	14
SE at LOSS	425	53	40	67
Math Content				
SE at HOSS	650	43	24	30
SE at Level 1/2	617	32	18	23
SE at Level 2/3	575	22	15	16
SE at Level 3/4	520	19	15	13
SE at Level 4/5	473	22	19	20
SE at LOSS	400	40	36	40
Math Process				
SE at HOSS	650	52	35	46
SE at Level 1/2	617	34	27	29
SE at Level 2/3	575	28	21	22
SE at Level 3/4	520	21	24	21
SE at Level 4/5	473	24	31	27
SE at LOSS	400	49	50	55
Science				
SE at HOSS	650	27	25	32
SE at Level 1/2	625	22	20	26
SE at Level 2/3	580	17	16	20
SE at Level 3/4	525	18	17	18
SE at Level 4/5	484	23	21	19
SE at LOSS	375	52	55	37
Social Studies				
SE at HOSS	625	27	29	23
SE at Level 1/2	619	24	29	23
SE at Level 2/3	580	19	24	19
SE at Level 3/4	529	18	31	18
SE at LOSS	400	42	38	34

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

Table 21. **Standard Errors at HOSS, LOSS and at each Proficiency Level Cut Score for each Cluster: Grade 8**

		Cluster		
Reading	**Scale Score**	**8A**	**8B**	**8C**
SE at HOSS	650	48	70	74
SE at Level 1/2	650	48	70	57
SE at Level 2/3	580	23	27	23
SE at Level 3/4	530	13	12	10
SE at Level 4/5	490	11	11	9
SE at LOSS	375	72	52	57
Writing				
SE at HOSS	625	56	52	70
SE at Level 2/3	551	25	31	35
SE at Level 3/4	505	24	29	23
SE at LOSS	425	35	35	30
Language Usage				
SE at HOSS	625	28	30	37
SE at Level 2/3	565	17	17	16
SE at Level 3/4	509	17	17	16
SE at Level 4/5	474	17	18	15
SE at LOSS	425	29	26	20
Math Content				
SE at HOSS	650	18	24	26
SE at Level 1/2	618	12	18	16
SE at Level 2/3	579	10	13	12
SE at Level 3/4	525	11	13	13
SE at Level 4/5	481	18	19	18
SE at LOSS	400	49	49	53
Math Process				
SE at HOSS	650	32	30	45
SE at Level 1/2	618	24	22	33
SE at Level 2/3	579	19	17	25
SE at Level 3/4	525	20	20	24
SE at Level 4/5	481	26	24	25
SE at LOSS	400	77	96	46
Science				
SE at HOSS	650	28	26	30
SE at Level 1/2	619	20	20	23
SE at Level 2/3	576	13	16	17
SE at Level 3/4	532	14	14	16
SE at Level 4/5	482	21	15	22
SE at LOSS	375	89	37	60
Social Studies				
SE at HOSS	650	31	30	37
SE at Level 1/2	620	24	22	26
SE at Level 2/3	582	17	16	21
SE at Level 3/4	530	15	15	14
SE at Level 4/5	495	17	17	14
SE at LOSS	375	50	46	49

Note: HOSS is the highest obtainable scale score, LOSS is the lowest obtainable scale score.

Table 22. Between Content Area Scale Score Correlations for Grade 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.62	1.00				
Lang. Usage	.63	.78	1.00			
Mathematics	.70	.63	.64	1.00		
Science	.76	.65	.65	.79	1.00	
Social Studies	.78	.67	.68	.77	.79	1.00

Note: N ranges from 57,132 to 62,009.

Table 23. Between Content Area Scale Score Correlations for Grade 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.55	1.00				
Lang. Usage	.60	.73	1.00			
Mathematics	.63	.59	.64	1.00		
Science	.69	.58	.63	.77	1.00	
Social Studies	.71	.60	.64	.71	.75	1.00

Note: N ranges from 55,667 to 60,141.

Table 24. Between Content Area Scale Score Correlations for Grade 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.68	1.00				
Lang. Usage	.71	.84	1.00			
Mathematics	.62	.61	.63	1.00		
Science	.74	.64	.67	.76	1.00	
Social Studies	.66	.63	.65	.69	.75	1.00

Note: N ranges from 52,982 to 56,396.

Table 25. Between Content Area Scale Score Correlations at School Level for Grade 3

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.94	1.00				
Lang. Usage	.93	.95	1.00			
Mathematics	.94	.93	.90	1.00		
Science	.96	.94	.92	.97	1.00	
Social Studies	.97	.95	.93	.96	.97	1.00

Note: N=801

Table 26. Between Content Area Scale Score Correlations At School Level for Grade 5

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.93	1.00				
Lang. Usage	.92	.95	1.00			
Mathematics	.91	.92	.92	1.00		
Science	.94	.93	.92	.97	1.00	
Social Studies	.94	.94	.92	.95	.97	1.00

Note: N=797

Table 27. Between Content Area Scale Score Correlations at School Level for Grade 8

	Reading	Writing	Language Usage	Mathematics	Science	Social Studies
Reading	1.00					
Writing	.94	1.00				
Lang. Usage	.94	.97	1.00			
Mathematics	.91	.91	.92	1.00		
Science	.96	.94	.94	.97	1.00	
Social Studies	.95	.95	.95	.94	.98	1.00

Note: N=252.

Table 28. Number of Items Flagged as Differential Item Functioning for 1997 MSPAP

Grade 3

	Reading (48 items)		Writing (11 items)		Language (28 items)		Math Content (80 items)		Math Process (46 items)		Social Studies (55 items)		Science (50 items)	
	+[1]	-[2]	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	2	2	0	0	1	0	2	2	2	1	1	1	1	0
Hispanic	0	0	0	0	0	0	0	0	2	1	0	0	1	2
Female	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Grade 5

	Reading (54 items)		Writing (11 items)		Language (31 items)		Math Content (68 items)		Math Process (31 items)		Social Studies (55 items)		Science (61 items)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	1	0	0	0	0	0	2	0	0	0	1	1	1	0
Hispanic	0	0	0	2	0	0	1	0	0	0	0	0	0	0
Female	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Grade 8

	Reading (59 items)		Writing (12 items)		Language (31 items)		Math Content (76 items)		Math Process (27 items)		Social Studies (54 items)		Science (61 items)	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asian	3	3	0	0	0	0	2	2	2	4	0	0	0	0
Hispanic	1	1	0	0	0	0	0	0	0	0	0	1	2	2
Female	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note 1: The minority group members did better than was expected
Note 2: The minority group members did less well than was expected

Table 29. Outcome Difficulty Indicators for each Grade for the 1997 MSPAP

Outcome Number	Outcome	Grade3	Grade5	Grade8
	Reading			
1.	Reading for Literary Experience	59	48	54
2.	Reading to be Informed	39	50	41
3.	Reading to Perform a Task	36	53	59
	Writing			
1.	Writing to Inform	34	37	48
2.	Writing to Persuade	35	39	44
3.	Writing to Express Personal Ideas	30	38	53
Language Usage				
Language In Usage		32	37	47
Mathematics				
Problem Solving		18	45	N/A
Communication		32	41	29
Reasoning		32	40	30
Connections		36	36	38
Concepts/Relationships		34	43	32
Measurement/Geometry		42	38	28
Statistics47		49	36	
Probability		42	42	36
Patterns/Relationships		44	N/A	N/A
Patterns/Algebra		N/A	44	24
Science				
Concepts of Science		39	47	39
Nature of Science		45	35	42
Habits of Mind		48	37	47
Processes of Science		33	40	40
Applications of Science		33	42	36
Social Studies				
Political Systems		40	49	45
People/Nation & World		34	33	43
Geography		41	50	39
Economics		31	37	47
Skills and Processes		36	42	45
Valuing Self and Others		34	43	40
Understand/Attitudes		32	49	42

Note: N/A means the outcome is not measure at that grade.
Note: The numbers are percentages of the maximum possible scores.

Appendix A

Test Maps for 1997 MSPAP

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1997 ADMINISTRATION

GRADE 3 FINAL TASK CALENDAR

	Tasks By Day Of Testing														
	MONDAY MAY 12			TUESDAY MAY 13			WEDNESDAY MAY 14			THURSDAY MAY 15			FRIDAY MAY 16		
	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times
A	3075 3008 3054	SCI SS M	45 35 25	3048 3064	R M/LWP	45 60	3060 3048	SS/LWP EWP SURVEY	45 40 20	3048 3063 3023	EWP M SS	55 25 30	3067	R/SCI	105
B	3076 3062	SS/LWP SCI	40 60	3065	R/LWP/ SCI/SS	105	3065 3007	R/SCI/SS SS	70 35	3066 3065	M EWP/SCI SURVEY	45 50 10	3065 3068	EWP M SURVEY	55 45 10
C	3074	R/LWP/SS SURVEY	95 10	3073	R/SS SURVEY	95 10	3073 3059 3072	EWP M SS	40 35 30	3073 3071	EWP SCI	55 55	3069 3070	LWP/M/ SCI M	70 35

▲ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

* Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1997 ADMINISTRATION

GRADE 5 FINAL TASK CALENDAR

	Tasks By Day Of Testing														
	MONDAY MAY 5			TUESDAY MAY 6			WEDNESDAY MAY 7			THURSDAY MAY 8			FRIDAY MAY 9		
	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times	#	Subject▲	Times
A	5059 5056	M M/LWP SURVEY	60 35 10	5057	R/SCI	105	5057 5071	EWP R/LWP	40 65	5057 5072	EWP SS SURVEY	55 45 10	5072 5073	SS SCI	50 55
B	5074	R/LWP/SS SURVEY	95 10	5062	M/SCI	105	5061 5046	SCI R SURVEY	50 50 10	5046 5069 5058	EWP M/LWP M	40 40 25	5046 5064	EWP SS	55 50
C	5067 5066	SCI/SS/ LWP M	80 30	5075 5020	SCI/R SS	75 30	5075	R/M/SCI	105	5075 5068	EWP SS/LWP SURVEY	40 55 10	5075 5070	EWP M SURVEY	55 40 10

▲ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

* Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

MARYLAND SCHOOL PERFORMANCE ASSESSMENT PROGRAM

DATES/TIMES* FOR MAY 1997 ADMINISTRATION

GRADE 8 FINAL TASK CALENDAR

	Tasks By Day Of Testing														
	MONDAY MAY 12			TUESDAY MAY 13			WEDNESDAY MAY 14			THURSDAY MAY 15			FRIDAY MAY 16		
	#	Subject^	Times	#	Subject^	Times	#	Subject^	Times	#	Subject^	Times	#	Subject^	Times
A	8067 8056	R/LWP SCI SURVEY	60 40 10	8050	R/SCI/SS	105	8050 8051	M SS/LWP	60 45	8050 8061 8049	EWP SS M	40 35 30	8050 8052	EWP M SURVEY	55 40 10
B	8055 8024	SS/LWP M SURVEY	40 40 20	8065	M/SCI/SS	105	8062 8065	R/LWP EWP	60 40	8065 8068 8053	EWP SS M	55 30 25	8063	R/SCI	105
C	8060 8064	M SS SURVEY	30 65 10	8066	R/SCI/SS	105	8066 8044	R/LWP/ SCI/SS R	55 50	8044 8057	EWP M/LWP	40 60	8044 8059	EWP SCI SURVEY	55 45 10

^ Language usage activities are distributed throughout and therefore not listed. Check your Examiner's Manual to determine where they occur.

* Each day is approximately 1 hour + 45 minutes of engaged testing and does not include time for organizing and preparing students for test administration.

BEST COPY AVAILABLE

Appendix B

Number of Items Comprising Each Outcome for 1997 MSPAP

Number of Measures for Each Outcome-Grade 3

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	6	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	7	8	11	4	9
Math Process	4	15	11	7	0	0	0	0	0
Social Studies	6	4	5	0	10	6	3	0	0
Science	5	5	4	0	5	5	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	6	0	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	9	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	5	8	5	5	9
Math Process	0	11	8	4	0	0	0	0	0
Social Studies	0	4	4	4	6	4	3	0	0
Science	9	4	5	0	4	7	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	6	6	0	0	0	0	0
Writing	2	1	0	0	0	0	0	0	0
Language Usage	9	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	12	5	6	6	9
Math Process	1	11	10	7	0	0	0	0	0
Social Studies	5	0	5	4	6	4	3	0	0
Science	4	6	5	0	7	5	0	0	0

Number of Measures for Each Outcome-Grade 5

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	6	0	0	0	0	0
Writing	2	0	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	3	6	4	0	4
Math Process	4	8	5	3	0	0	0	0	0
Social Studies	0	4	6	5	5	4	3	0	0
Science	4	4	4	0	6	4	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	7	6	0	0	0	0	0	0
Writing	1	1	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	9	5	4	4	5
Math Process	1	7	5	4	0	0	0	0	0
Social Studies	5	5	0	6	7	4	3	0	0
Science	6	5	3	0	6	7	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	6	6	0	0	0	0	0
Writing	1	2	0	0	0	0	0	0	0
Language Usage	9	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	6	4	6	6	5
Math Process	1	5	5	4	0	0	0	0	0
Social Studies	5	4	6	0	7	3	4	0	0
Science	8	7	6	0	8	7	0	0	0

Number of Measures for Each Outcome-Grade 8

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	0	6	6	0	0	0	0	0
Writing	2	0	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	7	6	5	6	13
Math Process	0	5	6	4	0	0	0	0	0
Social Studies	4	4	0	4	5	3	6	0	0
Science	5	5	5	0	5	4	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	6	0	7	0	0	0	0	0
Writing	0	2	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	4	4	5	4	6
Math Process	0	5	2	4	0	0	0	0	0
Social Studies	0	4	7	4	9	3	3	0	0
Science	6	8	6	0	8	5	0	0	0

	Outcome Number								
	1	2	3	4	5	6	7	8	9
Reading	0	8	6	0	0	0	0	0	0
Writing	2	0	1	0	0	0	0	0	0
Language Usage	8	0	0	0	0	0	0	0	0
Math Concept	0	0	0	0	5	6	6	4	6
Math Process	0	5	4	4	0	0	0	0	0
Social Studies	4	0	4	5	8	4	4	0	0
Science	5	3	4	0	4	7	0	0	0

Appendix C

Scaled Score Ranges for Each Proficiency Level in MSPAP`

MSPAP Proficiency level scale score ranges

Level	Grade 3	Grade 5	Grade 8
READING			
1	620-700	620-700	620-700
2	580-619	580-619	580-619
3	530-579	530-579	530-579
4	490-529	490-529	490-529
5	350-489	350-489	350-489
WRITING			
1	614-700	- - - -	- - - -
2	577-613	567-700	551-700
3	528-576	522-566	505-550
4	350-527	488-521	350-504
5	- - - -	350-487	- - - -
LANGUAGE USAGE			
1	620-700	597-700	- - - -
2	576-619	567-596	565-700
3	521-575	533-566	509-564
4	350-520	350-532	474-508
5	- - - -	- - - -	350-473
MATHEMATICS			
1	626-700	617-700	618-700
2	583-625	575-616	579-617
3	531-582	520-574	525-578
4	489-530	473-519	481-524
5	350-488	350-472	350-480
SCIENCE			
1	619-700	625-700	619-700
2	580-618	580-624	576-618
3	527-579	525-579	532-575
4	488-526	484-524	482-531
5	350-487	350-483	350-481
SOCIAL STUDIES			
1	622-700	619-700	620-700
2	580-621	580-618	582-619
3	525-579	529-579	530-581
4	495-524	350-528	495-529
5	350-494	- - - -	350-494

Dashes indicate proficiency levels for which cut scores could not be established for MSPAP. These cut scores will be established on future editions of MSPAP.